TOWARD AN EVALUATIVE METHODOLOGY FOR CRITERION-REFERENCED MEASURES: TEST SENSITIVITY

Dan Gilbert Ozenne

CSE Report No. 72
October, 1971

PROBE
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

This paper is based on a dissertation submitted in partial fulfillment of the requirements for the Doctor of Philosophy in Education, University of California, Los Angeles, 1971.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Theories of psychological measurement have grown out of the observation that individuals are different. The development of psychological tests came from an attempt to quantify these differences, and the result has been a prevailing concern with developing tests which are sensitive to these differences. A classic example of this concern can be seen in the work on aptitude measures by Binet.

In aptitude measurement one assumes that there is some underlying psychological trait which is important in some way. Measures of the trait are concerned with making distinctions among individuals with regard to how much of the trait in question each individual possesses. The important qualities of such a measure are that it is capable of making fairly fine distinctions and that it is stable over time, i.e., that distinctions among a given set of individuals will be similar at some time in the future.

In order to facilitate the understanding and interpretation of individual measurements they are usually referenced to the average levels of performance on the measure for some specified group. Thus, when one speaks of a child's IQ score, the score carries with it certain information about the child's performance relative to that of other children of his age.

The model of measurement resulting from this concern with individual differences has been used rather extensively. It has been used as a basis for the development of aptitude, attitude, and achievement measures. From it have come various indices and guidelines useful in test construction.

So prevailing is this model that the psychometrician begins rather automatically to think in terms of this methodology when asked to develop a test.

Yet, as prevalent as this approach is, it may not always be appropriate. There are instances when one is not concerned with an individual's performance relative to a group average, but rather with the absolute level of his performance. For example, if one were responsible for granting life guard certificates one would want to be assured that each individual surpasses some minimum level in the performance of each of the relevant skills.

The emphasis in the above situation has shifted from performance relative to other individuals to performance relative to some specified standard. While some may object to the above example, saying that these skills are motor and of not very great psychological interest, it is not too difficult to find examples of more psychological importance. For example, in any learning situation where there is a certain set of specified skills requisite to subsequent learning (or performance), a procedure of testing for minimum performance levels is meaningful.

The point to be made is that there exists some class of psychologically meaningful measures where the emphasis is on measuring the level of performance relative to some standard for each subject rather than the level of his performance relative to others. With the shift in emphasis in the measurement situation, traditional methods for evaluating the measure may be inappropriate. It is the purpose of this paper to first show that the two conceptualizations of the measurement situation do, in fact, lead to different concerns for evaluating the measure and, second, to introduce an alternative method for evaluating criterion-referenced measures. The developmental theory for this alternative

methodology, as well as some data relevant to a study of the usefulness of the method will be introduced and discussed.

Overview

In the previous section the distinction was made between the traditional, norm-referenced approach to measurement, and a newer approach based on the performance of an individual relative to some criterion. The research reported in the following chapters is mainly concerned with the development and evaluation of the latter class of measures.

The second chapter elaborates the basis for the distinction between the two types of measures and suggests that methods for evaluating one type of measure may be inappropriate for evaluating the other.

Chapter III provides a review of the literature relevant to the evaluation of the objective-based measures. At this point the concept of sensitivity is introduced as an appropriate method for evaluating such measures.

The traditional model for the response of a subject to a measure is presented in Chapter IV. It is shown how this model leads to an estimate of the reliability of the norm-referenced measure. This basic response model is then extended to conform to the typical objective-based measurement situation. At this point it is suggested how the response model can be used to evaluate the sensitivity of a measure.

In Chapter V the methods for assessing the sensitivity of a measure are more fully developed. Alternative versions of the response model are presented to account for a variety of measurement situations. The role of item selection and the effect of guessing in the evaluation of the measure are also introduced.

Chapters VI, VII and VIII introduce data from a variety of sources, both simulated and empirical, which were used to study the effects of varying test parameters on the sensitivity of the measure and which gave information regarding the results of the item analysis and selection techniques presented in Chapter V.

Finally, the implications of the results from the various data sources on the sensitivity of the measure are discussed in Chapter IX. Here some general considerations for test development are given, based on the results of the various empirical trials. Recommendations for further research regarding the proposed methodology are also included.

CHAPTER II

THE TWO APPROACHES TO MEASUREMENT

The previous chapter described two general approaches to measurement. Measures which yield information about a student's performance relative to the performance of others have been termed norm-referenced measures (NRM). Measures which yield information in terms of specific levels of performance, without reference to the performance of other subjects, have been called mastery tests, objective-based tests, or more popularly, criterion-referenced measures (CRM). These distinctions have been previously noted (Glaser, 1963; Coulsen and Cogswell, 1965; Ebel, 1966; Popham and Husek, 1969) and are widely used.

In the following chapters these two approaches will be presented in greater detail with an emphasis on those aspects of the underlying philosophies which lead to the development of the various methods for evaluating tests. It will be shown that the two approaches lead to different notions of desirable test characteristics.

Norm-Referenced Measures

The premise underlying the development of NRM is that individuals vary with respect to the amount each possesses of the psychological trait in question. Furthermore it is considered that a good test is one which maximally differentiates individuals' performances with respect to the trait. With this underlying philosophy it is not surprising to find that NRMs are constructed so as to maximize the discriminations made among individuals. The test construction and evaluation methodology is based on an attempt to obtain this maximization.

Typically, tests are evaluated in two ways; i.e., in terms of validity and reliability. A test is valid if it measures what it purports to measure. A test is reliable if its measurements are stable. There are various methods for investigating each of the evaluative aspects.

The APA Technical Recommendations for Psychological Tests and Diagnostic Techniques (1954) lists four types of validity: predictive, concurrent, construct, and content. The first three of these depend upon correlational data and are therefore dependent upon variability in the set of obtained scores. The fourth, content validity, relies only upon the judgement of so-called experts to determine if the test is really measuring the trait in question. It is interesting to note that the concept of content validity is directly applicable to all types of measurement, criterion-referenced measures as well as norm-referenced measures.

Reliability is usually ascertained by one of three methods; test-retest, parallel forms, or internal consistency. The first two methods are again correlational and are therefore dependent upon score variability. In test-retest reliability, for example, the scores on a test given on two occasions to the same subjects are correlated. Given that there is some error in psychological measures, it makes intuitive sense (as well as being mathematically demonstrable) that if subjects' scores are very close together on one occasion, small changes in these scores on the next occasion can lead to a different ordering of the individuals, thus suppressing the correlation. On the other hand, if scores are widely spread on one occasion, then small changes will not affect the relative order of the subjects and correlation will be high.

The third type of reliability is based upon the homogeneity of the set of items. Homogeneous tests, i.e., tests whose items all measure the same trait, maximize the likelihood of observing individual differences.

In practical test construction, the variability of the scores can be manipulated by item construction and selection techniques. Item content and difficulty are manipulated by expert item writers to obtain a test with maximal differentiating characteristics. For example, items that are so difficult (easy) that everyone fails (passes) are usually excluded from norm-referenced measures because they add nothing to the variability of the distribution of total scores. Such items may nevertheless reflect the appropriate content.

Another important aspect of test evaluation concerns the use to which a test is to be put. Binet's pioneering work in aptitude measurement was undertaken to identify the most feeble-minded students so that they could be placed in special schools with limited programs (Cronbach, 1960). The methodology that evolved from this early work has been described above. It is clear that if properly carried out, this methodology leads to measures which are effective in ranking subjects with respect to psychological traits (e.g., amount of mechanical aptitude, degree of depression, or empathy for minority groups). Thus in any situation where such rankings are needed, norm-referenced measures are appropriate. For example, norm-referenced measures are used in schools to assign grades.

Another instance where NRMs are appropriate is in the selection of a limited number of subjects for some subsequent treatment. Examples of this use include admissions to college where the admissions officer

has traditionally been concerned with each applicant's relative likelihood of success (Klein, 1970).

It should be noted that NRMs can also be used for comparing groups. For example, if a school's principal is interested in ascertaining the performance of his school's mathematics department he could compare his school's percentile to that of other schools with similar characteristics.

From the preceding it seems obvious that NRMs are appropriate whenever there is a need to order individuals' performances or compare an individual or group to other individuals or groups in terms of rank. What may not be so obvious is that there exists a large class of measurement situations where NRMs are not appropriate. An example which is of great educational importance is a situation where one wishes to ascertain the level of proficiency that an individual or a group has achieved.

Criterion-Referenced Measures

Garvin (1970) pointed out that "there are certain tasks that, by their very nature, must be performed at a specifiably high level in almost every imaginable situation." Among these are practically every task which involves public safety; for example, an examination of requisite skills for lifeguards. An example closer to academic interests might be the English examination used by many colleges for placing freshmen in either the regular English composition classes or in remedial, so-called "bonehead" English classes. The implicit assumptions in this class of measures are that there exists some set of skills necessary to later success, that these skills can be specified, and that they can be measured.

In the example of the English placement examination it is assumed that some set of skills is necessary to college success. Entering students are tested on their mastery of these skills. Those that are apparently lacking in the requisite skills are given remedial instruction designed to raise the students' proficiencies in these skills.

There are two important aspects to the example given above. First, there exists some criterion to which the test is referenced. This allows for the specification of the requisite skills and measurement of those skills. Second, one is not concerned with the test's ability to differentiate among individuals. In this situation it is irrelevant whether Joe has a higher score than Jack. Each student is compared only with the set of desired skills. If a subject possesses a sufficient number of skills he enters a regular English class, if not he enters a remedial class before going on.

The different emphasis in this class of measurement leads to a different approach to test development. Here no concern need be given to whether items discriminate among individuals. Individual differences and score variability have become irrelevant (Popham and Husek, 1969).

It would seem from the above that criterion-referenced measurement may be the appropriate approach whenever one needs to describe an individual's performance relative to some specified standard. In many educational settings, classes are of a cumulative or sequential nature where understanding of later content is dependent upon mastery of earlier content. In such situations, if one can specify the important content, criterion-referenced measures are appropriate.

If one accepts that there exists a class of measures where a CRM

approach is more appropriate than NRM, then one may ask how CRMs are to be evaluated. The methodology for evaluating NRMs has been established for some time and is detailed in numerous texts (Popham, 1970). Such is not the case for CRMs. In fact, a new textbook on measurement by Brown (1970) gives but one paragraph to CRM. Although Popham and Husek (1969) pointed out the inadequacies of traditional measurement theory for developing and analyzing CRMs, little has been done to provide an alternative methodology for this class of measures. That traditional evaluative aspects of measurement may not be applicable has been noted as recently as the 1970 AERA Symposia on Criterion-Referenced Measures by Cox (1970) and Popham (1970), both of whom discussed possible item analysis techniques.

Traditional methods of evaluating NRMs may be inappropriate for evaluating CRMs because these methods depend upon score variability (Popham and Husek, 1969). In an idealized situation where a criterion-referenced measure is given before and after an instructional unit one might find that subjects failed all of the items before instruction and passed all of the items after instruction. Certainly one could not fault such a test and yet under the norm-referenced methodology the items and the test must be considered worthless because there is no between-subject variability. The approach in this paper is an attempt to develop procedures to aid in the evaluation of CRMs that are consistent with the situations where such measures are appropriate.

CHAPTER III

THE PROBLEM IN THE CONTEXT OF RECENT DEVELOPMENTS

Norm-referenced measures are typically evaluated with regard to the constructs of reliability and validity. While it has been suggested that these constructs may not be directly applicable to CRM (Popham and Husek, 1969; Cox, 1970) an alternative evaluative methodology has not as yet been fully developed. Some notable work in this direction has, however, recently taken place.

It was noted earlier that criterion-referenced measures are based on the specification and measurement of subject skills. In instructional settings these specified skills may be stated as learning objectives. In order to insure the validity of the measure a relationship must be obtained between objectives and test items which will equate achievement on the test to achievement of the objective. This relationship has been the subject of recent research by Dahl (1971), who refers to the relation between objectives and items as "objective-item congruence." Content validity, as previously noted, plays an important role in this methodology. While the topic of validity will not be treated in this paper, the importance of the adequacy with which the test items measure the objectives can hardly be over emphasized. In all of the development which follows a permeating concern for validity, while not explicitly discussed, is implied.

Livingston (1971) has attempted to define a reliability coefficient for criterion-referenced measures. This methodology is based on defining variance about some criterion level as the variance of interest rather than the variance about the subject's mean. Because the variance about the mean is a minimum, Livingston's coefficient will always

be larger than the norm-referenced reliability estimate for the same data. The major criticism of this method is that the reliability defined in this way can easily be manipulated by changes in the criterion level. Since the criterion level is usually arbitrarily set, this is quite unfortunate because a researcher can easily raise his reliability by an arbitrary change in the criterion level.

Additional concern has been centered on item analysis techniques. Cox and Vargas (1966) introduced a discrimination index which they demonstrated leads to a somewhat different evaluation of test items than the traditional index based on discriminations by items between extreme groups. A comparison of several methods for evaluating items was undertaken by Popham (1970). An index for identifying atypical items in a set of comparable items was also suggested. While these item analysis and selection techniques are necessary tools for test construction (this will be discussed in more detail at a later point in this paper), they typically do not provide information about the adequacy of the test as a whole. Indeed, one could select a few of the best items from a pool of several items and still have a poor test if even the best items demonstrated only small instructional gains.

An attempt to use item analysis techniques to develop test evaluation indices has been undertaken by Ivens (1970). Ivens defines reliability indices based on the concept of within-subject equivalence of scores, i.e., item reliability is defined as the proportion of subjects whose item scores are the same on the posttest and either a retest or a parallel form. Score reliability is then defined as the average item reliability. Ivens also defines two indices of overall test effectiveness based on differences in performance levels on pretest, posttest, and retest. The need for retests or two forms of the posttest would

seem to reduce the usefulness of this methodology and limit its use to very special situations.

It is with the evaluation of the test as a whole that this paper is concerned. While Dahl's work provides some basis for assessing the validity of a criterion-referenced test, an adequate analog to the reliability construct has not been provided. It is felt that single indices of test usefulness, such as the reliability coefficient, should properly be derived from the philosophical conceptualization of the measurement situation. In NRM comparisons are to be made between individuals and, therefore, evaluative indices are based on the ability to make between-subject discriminations.

Concern in CRM centers not on comparisons between individuals, but rather on comparisons between groups of individuals who possess or do not possess the skill in question. The concern then for evaluating a given test is to determine how sensitive that test is to the presence of the relevant skills. In the pages which follow an index for test evaluation will be developed which is based on this concept of sensitivity. Item selection techniques will then be treated in their proper perspective; that is, as they contribute to the overall value of the test.

At this point it should be noted that a somewhat restrictive concept of CRM is to be used in the development that follows. Concern will be focused on the evaluation of cumulative or sequential instructional units as described in the English examination example (p. 8). The concern for measurement then becomes one of determining whether subjects possess the requisite skills for subsequent units. It is assumed that these skills or content knowledge can be specified. Such specific

skills or knowledge then become the objectives which the instrument is designed to measure.

The methodology for evaluating the instrument then becomes conceptually simple. All one need do is to compare the performances of those who possess the skills or knowledge with the performances of those who do not. In practice, however, it may be impossible to identify these two groups. Therefore, this development will restrict itself to situations where one can be reasonably certain that prior to a given instructional unit the level of knowledge of the content of that unit possessed by the subjects is quite low. Then, if one assumes instruction to be adequate, one can compare the performance on the measure after instruction with that prior to instruction. If the test is measuring the specified content, the differences in performance on these two occasions should be substantial. Therefore, concern here will be directed toward assessing the sensitivity of a measure to instruction. At a later point in this paper the assumption regarding the adequacy of instruction will be relaxed and various aspects of item evaluation under less than adequate instruction will be discussed.

CHAPTER IV

MODELS FOR SUBJECT RESPONSES

Reliability and Response Models

In attempting to measure the amount a subject possesses of a specified characteristic, psychometricians generally assume that the observed score (Y_{ij}) for an individual (j) on a measure (i) has two independent components. One is the magnitude, π_j, of the specific characteristic; the other is the error associated with the attempt to measure the characteristic (E_{ij}). The latter is due to both the measurement instrument itself and to the conditions surrounding the measurement situation. The general model then becomes:

$$Y_{ij} = \pi_j + E_{ij} \quad (1)$$

where Y_{ij} = observed measurement of person j on measure i

π_j = true magnitude of characteristic

E_{ij} = error in measurement

In norm-referenced measurement studies this model is used to assess the reliability of the test, i.e., the ability of the test to spread the individuals out. In order to determine the reliability, the variance of the population of scores is partitioned into true (between subject) and error (within subject) variance components. Subtracting π (the population mean) from both sides of Equation (1), squaring both sides and taking the expectation over the population of subjects while noting that the cross-product terms vanish on the right (since the components are independent), yields:

$$E(Y_{ij} - \pi)^2 = E(\pi_j - \pi)^2 + E(E_{ij}^2) \quad (2)$$

or

$$\sigma_y^2 = \sigma_s^2 + \sigma_e^2 \quad (3)$$

Reliability for norm-referenced measures is defined as that proportion of the variance of scores which is true variance:

$$(4) \qquad \rho_{11} = \frac{\sigma_s^2}{\sigma_y^2} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}$$

This formulation lends itself neatly to a partition of variance in the analysis of variance model. The model ordinarily used is the one-way analysis of variance with subjects as the factor of interest. Items are considered to be replications within subjects. The following table shows the analysis of variance breakdown with the variance components each source estimates. Here n subjects each respond to a items.

Table 1

Analysis of Variance for Between Subjects Design

Source	Degrees of Freedom	Expected Mean Square
Between subjects	n-1	$\sigma_e^2 + a\sigma_s^2$
Within subjects	n(a-1)	σ_e^2

The mean squares from such an analysis allow estimation of the necessary variance components needed to estimate the reliability of scores in the population. The methodology for such estimation is detailed in Winer (1962) and in Meyers (1966).

An Extended Response Model

If, however, one is interested in assessing the sensitivity of a criterion-referenced test to instruction, a different model is needed. It should be noted that the model to be presented is not a different conceptual model in terms of the representation of a subject's score but is rather an extension of the above model to account for score variability due to instruction.

A restatement of equation (1) in terms of a subject's deviation from the population parameter yields:

$$(5) \qquad Y_{ij} = \pi + \alpha_j + E_{ij}$$

where α_j is the deviation from the population parameter $(\pi_j - \pi)$. To this model can be added a dimension for time. Interest centers on but two occasions for the time variable; namely before and after instruction. The implicit assumption is that if there is any difference in level of responses on the two occasions that such a difference is due to the intervening instruction. An alternative model for investigating the effect of instruction will be presented at a later point in this paper.

If the time variable is added to the model in the form of deviations from the population parameter $\beta_k = \pi_k - \pi$ (now for two occassions) the model becomes

$$(6) \qquad Y_{ijk} = \pi + \alpha_j + \beta_k + E_{ijk}$$

The model presented here is basically that of an additive, subjects-by-occasions analysis of variance as presented in Meyers (1966, p. 154). In this model the α_j can be thought of as enduring individual differences and β_k as the effect of having or not having instruction.

The model above could be used to partition score variability into variance components. However, more often than not (especially when time periods are arbitrarily fixed) the variability among subjects' scores will be a function of the particular occasion under observation. This means that an interaction of subject and occasion level contributes to the score. If such a situation exists the above model should be revised to include an interaction term in the population as a contribution to the score Y_{ijk}. This is called a non-additive model (Meyer, 1966) and is represented as

$$(7) \qquad Y_{ijk} = \pi + \alpha_j + \beta_k + (\alpha\beta)_{jk} + E_{ijk}$$

The above model would seem to be complete in accounting for the variability of scores on a set of comparable measures (usually items) administered before and after instruction. This model will be used in the following development.

In contrast to the norm-referenced measurement model where interest lies with the between subjects variability, the interest with this model for criterion-referenced measures will lie in the between occasions variability. The model, as presented here, is in agreement with Popham and Husek's (1969) conceptualization of subject variability as an irrelevant dimension. This model still allows for individual variability but considers such variability to be irrelevant to the purpose of such studies, namely, to assess the measure's sensitivity to instruction. As expressed in the present model such sensitivity would be manifested as a large occasions effect. If this model is used in the partitioning of score variability, a comparison can be made between the occasions variance and variance due to the error of measurement in a manner analogous to that used to assess the reliability in norm-referenced

measures. In NRM a test is considered reliable if the between subjects variability is large relative to the error of measurement. Analogously, a test may be considered sensitive to the effects of instruction if the occasions variance is large relative to the error variance.

In the following chapter an estimation of the sensitivity index will be developed.

CHAPTER V

METHODS FOR ESTIMATING SENSITIVITY

Toward an Index of Sensitivity

Upon repeated measurement with comparable instruments, usually items, the parameters α_j, β_k, and $(\alpha\beta)_{jk}$ are assumed to remain constant whereas the E_{ijk} are assumed to vary. The mean of $\underline{n}$ such measures for each occasion may be represented as

$$(8) \qquad \frac{Y_{ijk}}{N} = \bar{Y}_{\cdot jk} = \pi + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \bar{E}_{\cdot jk}$$

The data matrix in Table 2 shows a representation of the data collection scheme. It is assumed that the $\underline{n}$ comparable measures are a random sample from some universe of such measures and are administered to $\underline{b}$ subjects on each of $\underline{a}$ occasions (a=2). If the α_j, β_k, and $(\alpha\beta)_{jk}$ remain constant for such measurement situations, the variance within person $\underline{j}$ on occasion $\underline{k}$ is considered to be due to the error of measurement. The variance of the subject means, on the other hand, is in part due to individual differences, instructional effects, and interaction, and in part due to differences in the average error of measurement for each subject.

Table 3 represents the breakdown of data from Table 2 according to the analysis of variance of Equation (7). It is assumed that both items and subjects represent random effects since they are considered to be random samples from their respective populations. The occasions effect, however, is considered fixed in that the two levels are arbitrarily selected and exhaust the levels of interest. The resulting expected mean squares are shown on the right side of Table 3. The symbol θ^2 is associated with the variance of a fixed effect, while σ^2 is associated with the variance of a random effect.

TABLE 2

Data Matrix

	Occasion 1			Occasion 2			
	Measures		Mean	Measures		Mean	
Person	1 . . i . . n			1 . . i . . n			Mean
1	Y_{111}	Y_{n11}	$\bar{Y}_{.11}$	Y_{112}	Y_{n12}	$\bar{Y}_{.12}$	$\bar{Y}_{.1.}$
.							
.							
j	Y_{ij1}			Y_{ij2}			
.							
.							
b	Y_{1b1}	Y_{nb1}	$\bar{Y}_{.b1}$	Y_{1b2}	Y_{nb2}	$\bar{Y}_{.b2}$	$\bar{Y}_{.b.}$
Means			$\bar{Y}_{..1}$			$\bar{Y}_{..2}$	$\bar{Y}_{...}$

TABLE 3

Analysis of Variance for Subject by Occasions Design

Source	df	EMS
A (occasions)	a-1	$\sigma_e^2 + bn\theta_a^2 + n\sigma_{ab}^2$
B (subjects)	b-1	$\sigma_e^2 + an\sigma_{ab}^2$
AB (interaction)	(a-1) (b-1)	$\sigma_e^2 + n\sigma_{ab}^2$
Error (within AB)	ab(n-1)	σ_e^2

Here $MS_{occasions}$ is defined as

$$MS_o = \frac{nb\sum_k^2(\bar{Y}..._k - \bar{Y}...)^2}{a - 1}$$

whereas the occasions variance for the subject means is given by

$$S_{\bar{s}}^2 = \frac{b\sum_k^2(\bar{Y}..._k - \bar{Y}...)^2}{a - 1}$$

where $S_{\bar{s}}^2$ signifies that the scores used are subject means. Thus

$MS_o = nS_{\bar{s}}^2$.

In terms of Table 3, the expected value of variance of the occasions for subjects' means is

$$E(S_{\bar{s}}^2) = b\theta_A^2 + \sigma_{AB}^2 + \sigma_{\bar{E}}^2$$

The quantity θ_A^2 is the variance of the occasions effect.

From the relationship between MS_o and $S^2_{\bar{s}}$,

$$E(MS_o) = nb\theta_A^2 + n\sigma_{AB}^2 + n\sigma_{\bar{E}}^2$$

$$= nb\theta_A^2 + n\sigma_{AB}^2 + \sigma_e^2$$

Thus: $n\sigma_{\bar{e}}^2 = \sigma_e^2$.

At this point the sensitivity of the test will be defined to be

$$(9)\ \mathcal{E} = \frac{\theta_A^2}{\theta_A^2 + \sigma_{\bar{e}}^2} = \frac{\theta_A^2}{\theta_A^2 + \frac{1}{n}\sigma_e^2}$$

In words, the sensitivity of a group of comparable measures given to a sample of subjects before and after instruction is the variance due to the instructional effect divided by the sum of the variance due to the instructional effect and variance due to the error of measurement.

Finally, all that is needed to estimate the sensitivity index are estimates of the values θ_A^2 and σ_e^2 . An examination of Table 3 shows that such estimates are available. The MS_{error} directly estimates σ_e^2 such that

$$(10)\ \hat{\sigma}_e^2 = MS_{error}$$

and

$$(11)\ \hat{\sigma}_{\bar{e}}^2 = \frac{MS_{error}}{n}$$

An estimate of θ_A^2 can be obtained from the $MS_{occasions}$ term and the $MS_{interaction}$ term as follows

$$(12)\ \hat{\theta}_A^2 = \frac{MS_{occasions} - MS_{interaction}}{Nb}$$

Thus the estimated sensitivity, upon substitution into equation (9), becomes:

$$(13) \qquad \hat{\mathcal{E}} = \frac{\hat{\theta}_A^2}{\hat{\theta}_A^2 + \hat{\sigma}_{\bar{E}}^2}$$

$$= \frac{\frac{1}{nb}(MS_{occas.} - MS_{inter.})}{\frac{1}{nb}(MS_{occas.} - MS_{inter.}) + \frac{1}{n} MS_{error}}$$

$$(14) \qquad = \frac{MS_{occasions} - MS_{interaction}}{MS_{occas.} - MS_{inter.} + bMS_{error}}$$

It should be obvious from the above formulation that as the occasions variance becomes large relative to the error of measurement that the sensitivity index will approach 1.0. Conversely if there is no occasions variance, i.e., if there is no instructional effect, the sensitivity index will go to zero.

An Alternative Model for Test Sensitivity

Just as there are alternative methods for estimating test reliability (e.g. test-retest, parallel forms) for NRM, there is an alternative method for assessing the sensitivity to instruction for a CRM. Rather than measuring the same individuals before and after instruction, one could measure two sets of persons who are similar except that one group has had the benefit of instruction while the other has not. Again it should be noted that in order to assess the sensitivity of a test to a set of objectives it is not necessary to give instruction on these objectives. But since it is ordinarily difficult to identify those who

possess these skills, in this development it is assumed that a group recently subjected to instruction can be considered as a group possessing the skills of interest. Therefore, once again a comparison will be made between a group subjected to instruction and a group without benefit of instruction. In any situation where a group that possesses the skills of interest can be identified a parallel method to that developed here can be used.

In this alternative model differences between individuals within either of the two treatment conditions are still seen as enduring differences in subject characteristics. Variability within a subject's responses to the comparable measures is again seen as being due to error associated with the attempt to measure the characteristic. The descriptive model for measurements taken under these conditions then becomes

$$Y_{ijk} = \pi + \beta_k + \alpha_{j/k} + E_{ijk} \tag{15}$$

where Y_{ijk} = observed response on measurement $\underline{i}$ for the $\underline{j}$th person under treatment $\underline{k}$

π = population parameter

β_k = effect of treatment

$\alpha_{j/k}$ = effect of the $\underline{j}$th subject in treatment $\underline{k}$

E_{ijk} = error of measurement

An analysis of variance framework can again be used to describe the various sources of variation in a subject's response. The appropriate model is that of a nested design with measures nested within subjects which in turn are nested within treatments. The analysis of

variance breakdown is shown in Table 4 along with the variance components associated with each source.

TABLE 4

Analysis of Variance for Nested Design

Source	Degrees of freedom	EMS
A (treatments)	a-1	$\sigma_e^2 + n\sigma_{B/A}^2 + nb\sigma_A^2$
B/A (subjects within treatments)	a(b-1)	$\sigma_e^2 + n\sigma_{B/A}^2$
Error (measures within subjects within treatments)	ab(n-1)	σ_e^2

Again interest generally centers on the mean (or total) performance of an individual over a series of comparable measures. In a manner analogous to that presented earlier in this paper it can be shown that $n\sigma_{\bar{e}}^2 = \sigma_e^2$ and, therefore, the sensitivity can be estimated if the treatment variance, σ_A^2, and the error variance, σ_e^2, can be estimated.

From an inspection of Table 4 it is clear that an estimate of the treatments variance is available from some manipulations of the between treatments mean square and the subjects within treatments mean square. This estimate is stated as:

$$\hat{\theta}_A^2 = \frac{MS_{\text{treatments}} - MS_{\text{subjects/treatments}}}{nb} \tag{16}$$

Again the error of measurement is directly estimable from the mean square error such that

$$(17) \qquad \hat{\sigma}^2_e = MS_{error}$$

and

$$(18) \qquad \hat{\sigma}^2_{\bar{e}} = \frac{MS_{error}}{n}$$

Thus the estimated sensitivity of the test to instruction, upon substitution into equation (13), becomes

$$\hat{\xi} = \frac{\frac{1}{nb}\,(MS_{treat.} - MS_{subs/treat})}{\frac{1}{nb}\,(MS_{treat.} - MS_{subs/treat}) + \frac{1}{n} MS_{error}}$$

$$(19) \qquad \hat{\xi} = \frac{(MS_{treat} - MS_{subs/treat})}{MS_{treat} - MS_{sub/treat} + bMS_{error}}$$

Accounting for Objectives

The two models thus far presented assume that all items measure the same objective. If it is desirable to measure competence on more than one objective, then one would want a model which takes differences between objectives into account. Items measuring different objectives would not necessarily be homogeneous and therefore the use of the previous models may result in an increased error variance and, consequently, a decrease in sensitivity.

In developing a model which takes differences in learning objectives into account, one encounters a minor philosophical problem. Some may contend that if items measure different objectives, then the proper procedure is to consider each such set as a separate test. It is the contention here that this problem is a psuedo-problem. One can write

objectives at almost any level of specificity that one desires. Certainly a logical, if somewhat extreme, case could be made for considering each item as a distinct entity, constituting a complete test of an extremely specific objective. On the other hand, some commonly used measures, such as an eighth-grade mathematics test, can be thought of as measuring a rather general objective.

The position to be taken here is that if one wishes to measure a rather broad objective, knowing that the items can be grouped by sub-objectives, then the proper procedure is to use a model which controls for variability due to the presence of the sub-objectives. The important point is that the sub-objectives are related and are subsumed under a higher-order, more general objective.

Within the analysis of variance framework previously presented the heterogeneity of items due to differences in objectives can readily be accounted for by introducing objectives as an additional factor. Since the objectives are purposely selected to reflect the goals of a particular instructional unit the objectives factor is seen as a fixed design factor. The appropriate linear model for a response is now

$$(20) \qquad Y_{ijkl} = \pi + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl} + E_{ijkl}$$

where

Y_{ijkl} = observed response on measure $\underline{i}$ for person $\underline{j}$ on occasion $\underline{k}$ and objective $\underline{l}$.

π = population parameter

α_j = individual differences

β_k = effect of occasion $\underline{k}$

γ_l = effect of objective $\underline{l}$

$(\alpha\beta)_{jk}$, $(\alpha\gamma)_{jl}$, = interaction effects
$(\beta\gamma)_{kl}$, $(\alpha\beta\gamma)_{jkl}$

ε_{ijkl} = error of measurement

Once again an analysis of variance framework is used to describe the different sources of variability in subjects' responses. The analysis of variance table shown below (Table 5) shows the sources of variability in the responses of a subjects to n items measuring each of c objectives on each of b occasions.

TABLE 5

Analysis of Variance for Three Factor Crossed Design

Source	Degrees of Freedom	EMS
A (subjects)	a-1	$\sigma_e^2 + nbc\theta_A^2$
B (occasions)	b-1	$\sigma_e^2 + nac\theta_B^2 + nc\sigma_{AB}^2$
C (objectives)	c-1	$\sigma_e^2 + nab\theta_C^2 + nb\sigma_{AC}^2$
AB	(a-1)(b-1)	$\sigma_e^2 + nc\sigma_{AB}^2$
AC	(a-1)(c-1)	$\sigma_e^2 + nb\sigma_{AC}^2$
BC	(b-1)(c-1)	$\sigma_e^2 + na\theta_{BC}^2 + n\sigma_{ABC}^2$
ABC	(a-1)(b-1)(c-1)	σ_e^2
Error	abc(n-1)	

The estimates of variance components needed for the sensitivity index are readily estimable from this breakdown. Again estimates of the occasions variance θ_A^2, and the error variance, σ_e^2, are needed. These are obtained as follows

$$\hat{\sigma}_e^2 = MS_{error}$$

$$\hat{\theta}_B^2 = \frac{1}{nac}\left(MS_{occas.} - MS_{subjects\ X\ occas.}\right)$$

The introduction of the objectives effect in the model can lead to a more accurate estimation of the error of measurement if in fact there are differences between subjects' performances on the various objectives. It would seem to be reasonable to expect such differences in many situations. For example, the concepts that each objective represents may vary in difficulty to learn. Most would agree that division is a more difficult concept to learn than addition, and yet both may appear as sub-objectives in a mathematics achievement test.

The second model presented, that for separate competency groups, can also be modified to take differences in objectives into account. To the linear model presented (Equation 15) three components are added. These are

γ_l = the effect of objective $\underline{l}$

$(\beta\gamma)_{kl}$ = the interaction of conditions and objectives

$(\gamma\alpha)_{lj/k}$ = the interaction of objectives and subjects within conditions

The analysis of variance framework outlined in Table 6 is used to estimate the needed variance components. These components, the variance between conditions and the error variance, are estimated by

$$\hat{\theta}^2_A = \frac{1}{nbc}(MS_{conditions} - MS_{subjects\ within\ conditions})$$

$$\hat{\sigma}^2_e = MS_{error}$$

TABLE 6

Analysis of Variance for Two Crossed and One Nested Factor Design

Source	Degrees of Freedom	EMS
A (conditions)	a-1	$\sigma^2_e + nbc\theta^2_A + nc\sigma^2_{B/C}$
B/A (subjects within conditions)	a(b-1)	$\sigma^2_e + nc\sigma^2_{B/C}$
C (objectives)	c-1	$\sigma^2_e + nab\theta^2_c + n\sigma^2_{CB/A}$
AC	(a-1)(c-1)	$\sigma^2_e + nb\theta^2_{AC} + n\sigma^2_{CB/A}$
CB/A	a(b-1)(c-1)	$\sigma^2_e + n\sigma^2_{CB/A}$
Error	abc(n-1)	σ^2_e

Item Selection

The approach thus far presented has emphasized differentiations between two instructional groups. These two groups have been designated as (a) a group of persons who are highly experienced and competent in the area of criterion performance and (b) the same group before they received any instruction in the content area. An alternative approach suggested that group "b" might also be a group similar to group "a" in all respects except instruction in the content area. An evaluative index was developed which was based upon differences in the responses of the two groups to a set of items. At this point it is necessary to consider the contribution of the individual items to the sensitivity of the test as a whole.

In the construction of NRM, items are selected which maximize the variability between subjects. But with CRM the main concern should be with maximizing the differences between the two competency groups. That these two concerns will lead to a somewhat different selection of items has been demonstrated empirically by Cox and Vargas (1966). These authors suggest that the appropriate index is one based on the differences in the percentage of students passing the item at pretest and posttest. Such an index would provide information about the items ability to discriminate between pre- and posttest performances.

In the ideal case an item would be failed by all subjects at pretest and passed by all subjects at posttest. Such an item is maximally sensitive to the instructional situation, demonstrating both a need for instruction when failed at pretest and the effectiveness of the instruction when passed at posttest. Short of the ideal case it is obvious that if one wishes to maximize the variance of responses between these two occasions then one should choose items which have the greatest amount of difference in performance levels for the two administrations. Therefore the discrimination index presented by Cox and Vargas would seem to be the most appropriate one for use with CRM.

Since Cox and Vargas worked only with the differences between the pretest and posttest performance of the same group of subjects, their index is expressed as the difference between the proportion of passes on these two occasions. For present purposes, and with no loss in meaning, a generalization of this technique will be defined. Item discrimination is here defined as the difference in the proportion of passes in the high competency group and the proportion of passes in the low competency group. In this manner the index is applicable to any of the

situations presented in this paper.

An alternative method of defining the CRM discrimination index has been suggested by Ivens (1970) when retest data is available. Due to the effort and difficulty involved in obtaining these data, however, it is not expected that the method will gain much use.

In addition to ranking items by their contribution to the discrimination between the two competency groups, one would also like to identify atypical items. In NRM an item with a negative item discrimination index or a very low positive index value is usually deleted from the test. The same can be done in CRM using the CRM discrimination index defined above. In addition, items which have undesirable characteristics (i.e., high proportions of pass-pass or fail-fail responses) should be inspected.

In an attempt to develop an index to identify poor items Popham (1970) has suggested a method based on the four possible outcome patterns for an item administered on two occasions (i.e., fail-fail, fail-pass, pass-fail, pass-pass). For each item one begins by tabulating the frequencies of each outcome category over all subjects. Popham then suggests that a "prototypic item" can be defined by taking the median frequency of each outcome category over all items. Each of the individual items can then be compared with the prototypic item on the basis of the frequencies in each of the four categories. The suggested method of comparison is to compute chi-square values for each item in comparison to the prototypic item. Large chi-square values would indicate that the response category frequencies for the item are considerably different from that for the typical item. Popham presents empirical data which would seem to support the usefulness of this approach

although appropriate limits of the chi-square value for item exclusion have not been investigated.

These two methods, the CRM item discrimination and the chi-square for atypical items, would seem to be a sound approach to the selection and analysis of items for a highly sensitive CRM. The usefulness of these techniques will be investigated as they apply to the empirical data presented in Chapters VII and VIII.

Guessing

In multiple-choice tests there may be an increase in error variance as a result of the subject's guessing. This will be particularly true at pretest when the overall level of knowledge is expected to be quite low. The problem of guessing and a possible procedure for handling this problem are presented here.

For illustrative purposes the following example may be useful. Consider a twenty-item test where all subjects know essentially nothing about the material at pretest and can answer eighteen of the items correct as a result of instruction at posttest. What would the effects of guessing be in such a situation? First, it seems obvious that the most immediate effect would be the increase in the observed scores if the above conditions reflect the true situation. For example, if the subjects guess on all of the pretest items and these are the usual four-part multiple-choice items, the net effect is that an observed mean of five is to be expected. The posttest mean, because of the decreased number of items on which guessing is possible, will be raised by only one-half an item.

The effect on the evaluative model previously presented is a reduction in the occasions variance, and, hence, a lowering of the

sensitivity index. Additionally, one could expect guessing to increase the within cells variance for the pretest scores. This follows as a result of the relationship between the mean and variance of the binomial distribution. The effect on the evaluative model is an increase in the error variability and again, a reduction in the sensitivity of the measure.

It is felt that the effects of guessing are a major factor only at pretest. If knowledge increases as function of instruction, guessing will have less and less of an effect. It is assumed here that after instruction the guessing effect is so minimal as to make the effort involved in correcting for its effect unwarranted.

There are alternatives for attempting to control the effects of guessing. First, one can use items that require subjects to furnish answers rather than select from a given set of alternatives. Given the popularity of the multiple-choice test, this alternative is probably the least attractive, although potentially the most appropriate. Second, one could use formulas to correct for guessing. These procedures are summarized in most measurement texts (for example, Nunnally, 1967). The reasearch on the effects of guessing and of the various corrections for guessing have been summarized by Price (1964). The method suggested here is a correction for guessing of pretest scores by the formula

$$R_c = R - \frac{W}{A - 1} \tag{21}$$

where

R_c = an estimate of the persons correct score

R = number of correct responses

W = number of incorrect responses

A = number of alternatives for each item

When applied to the number of items attempted by a subject at pretest the above correction should yield a more accurate estimation of the test's actual sensitivity.

CHAPTER VI

METHODS AND RESULTS: OVERVIEW

While analytically the method proposed here for evaluating criterion-referenced measures would appear to be a useful one, its value cannot be fully realized in the absence of supporting empirical data. The concept of sensitivity has a certain amount of theoretical appeal but one would surely doubt its usefulness if carefully constructed measures, used appropriately, produced only very low values. For this reason it was decided to use a variety of data sources to investigate the sensitivity concept under various conditions. The following two chapters report the methods and results used in these analyses.

The preliminary study was undertaken to assess the importance of various test parameters in determining the sensitivity of the measure. In order to be able to exercise a certain degree of control over the values of these parameters and to get a rather wide range of values, simulation data were used in this preliminary phase. The method and results of this study are reported in Chapter VII.

A variety of empirical data was gathered in order to investigate the role of sensitivity for different types of test data. Three sources were used, representing data gathered from (1) a graduate course in statistics, (2) a junior high school mathematics program evaluation, and (3) instructional units in phonics and geometry at the primary grades level. The data from the graduate statistics course represent a conscious attempt to develop a good criterion-referenced measure. One would expect the resulting test to demonstrate considerable sensitivity. The junior high school mathematics data represent the application of a

traditionally constructed measure before and after an instructional period. Even though the test purports to measure the content area under study, no objectives were specified and it is felt that this lack of a specific plan for the instructional unit (and, therefore, the measure) will result in an insensitive measure. The source is to be used here to demonstrate the use of item selection procedures. The third source is included because the two tests represented here (phonics and geometry) were teacher-made tests which were written to measure the specified objectives of their respective instructional units. These tests are of interest because they represent what can be done without benefit of item refinement.

Each of the above sources, aside from representing differing data sources for the analysis, also provide unique situations where the various methods for item analysis, correction for guessing, and accounting for objectives can be tried and compared.

The form of report for the different data sources will differ somewhat from traditional formats in that for each source a description of the data and the methods for analyzing those data will be immediately followed by the results of the analysis and some conclusions based on the analysis. In the final chapter these separate conclusions will be summarized and their interrelationships discussed.

CHAPTER VII

METHODS AND RESULTS: SIMULATION STUDY

Methods

In order to investigate the characteristics of the proposed sensitivity index under a variety of conditions, test data with varying characteristics were simulated and then analyzed. By using simulated data, the investigator has the ability to examine characteristics of the sensitivity estimate under a broad spectrum of conditions. In this study it was felt that the simulation method would allow a more complete investigation of the important attributes of the index than might be available using empirical data alone.

The first step in the simulation methodology involved determining what characteristics of the simulation data would be under the investigator's control. The characteristics that seemed obviously important were the parameters of the distributions of responses on each of the two occasions, i.e., the mean and variance. Additionally, it seemed important to allow for correlated responses over the two occasions since individual differences do exist and should be expected to persist over the instructional period. These five variables (i.e., means, variances, and correlation) then became the basis of the simulation effort. By simulating test data with differing values of these variables one could investigate the relative importance of each in determining the sensitivity of a measure.

A multivariate data generator computer program[1] was rewritten to produce data with the desired characteristics for any number of items and subjects specified by the user. This program includes the use of

[1]A basic multivariate data generator program was supplied by J. W. Keesling, University of California, Los Angeles.

random deviates so that a large number of replications of data with the same input parameters can be generated. Because of the random component, the resulting distributions of generated data will be distributed around the values of the input parameters.

To the data generator was added an analysis of variance program. Thus, for any given set of input parameters, data were generated, summary tables of input and output parameters printed, and the data analyzed, giving estimates of the variance components and sensitivity index.

Although a few initial trials with varying numbers of subjects and items were undertaken to define the relative importance of these attributes, it was felt that the most important aspect of this phase of the research was to investigate the role played by distributional parameters in determining test sensitivity. Toward this end the number of items and subjects were fixed in the simulations reported here. Thus, only changes in the distributional parameters would cause changes in test sensitivity. In this way the effect of such changes could be analyzed.

Data were then generated by systematically varying the parameters. In particular, means were varied to give score distributions reflecting both large change and no change. Because of the random component the latter occasionally resulted in an observed decrement in performance. This is a plausible, though perhaps not likely, outcome and therefore these data were retained.

The variability for each of the two score distributions was also manipulated. By manipulating the variability of each distribution separately one can investigate not only the effects of large or small variances, but the effect of heterogeneity of variance as well. This again seemed important since in real test results one could quite reasonably

expect very low performance with low variability at pretest and higher performance with greater variability at posttest. The opposite alternative is also completely plausible and was therefore included.

The ratio of the two standard deviations was used to define a new variable for later analysis. If this new variable could be found to be related to test sensitivity this would indicate that the latter index is related to the homogeneity of test variances on the two occasions. Such a result would certainly limit the applicability of this technique of test evaluation.

Finally, various degrees of correlation of subjects' responses between the two occasions were produced. A high correlation coefficient indicates a strict preservation of individual differences across the instructional period; that is, the ordering of individuals would be highly similar on each of the two occasions. While in certain situations one might expect such individual differences to exist, in terms of conceptualization and model presented here they are irrelevant. Therefore, no relationship between the correlation between subjects' reponses on different occasions and test sensitivity is to be expected.

A total of 535 separate sets of test data were generated with varying parameters. While a wide range of values for each parameter was generated, particular emphasis was placed on generating values approaching what might be considered a good test. That is, the data reflect some concentration on producing data which show an increase in performance from pre- to posttest. The following section describes the resulting data and analysis.

Results

The data resulting from the 535 simulations of a 10-item test given to 20 subjects are summarized in Table 7. Here the mean, maximum and minimum values for each of the five test parameters are given. In addition, two new variables are defined which are derived from these parameters. First, the difference score is defined as the amount of change between pre- and posttest for each of the simulations. The variability ratio is defined as the ratio of the pretest standard deviation to the posttest standard deviation.

The results indicate that a rather broad range for each parameter was successfully obtained. Moreover these data would seem to reflect a realistic range of expected outcomes for a measure in which positive change is anticipated. The range of the difference scores reflect tests with a large increase in level of performance as well as tests which show a small decrement. The standard deviations reflect small to large variabilities in the distributions for the separate occasions. The ratio of the standard deviations indicate large heterogeneity in the extremes with pretest standard deviation roughly one-fourth as large as posttest standard deviation in one extreme and roughly four times as large in the other extreme.

The actual range of the computed value of the index indicates that the data represent both very good and very poor tests. The negative value indicated here is an artifact of the use of analysis of variance techniques for estimating variance components. Under this methodology one will occasionally obtain negative variance estimates. Thompson (1962) suggests that the best estimate of the true variance in such situations is zero. Since the negative variance component in these data is always the occasions variance, one may assume that the best estimate of the true occasions variance is zero and, therefore, the sensitivity value is zero.

TABLE 7

Summary of Simulated Test Characteristics
and Their Relation to Sensitivity

Characteristic	Mean	Max.	Min.	Correlation with index
Pretest Mean	2.89	5.10	.90	-.84 *
Standard Deviation	1.62	2.75	.77	-.23 *
Posttest Mean	6.46	8.85	3.70	.93 *
Standard Deviation	1.89	3.29	.66	-.18 *
Correlation	.36	.80	-.15	-.05
Difference	3.57	7.80	-.75	.94 *
Variance Ratio	1.01	3.80	.28	-.03
Sensitivity Index	.60	.97	-.27	

N = 535 *significant at p .01

Of particular interest in Table 7 is the last column on the right. Here is indicated the linear correlation of each of the variables with the sensitivity value. While linear relationships may not adequately describe the actual relationships between the variables, they can provide important clues to these relationships. Here it is immediately obvious that the most important characteristic of the testing situation is the amount of change which occurs. This change is dependent upon both the level of pretest performance and the level of posttest performance. Clearly the best of all possible criterion-referenced measures is that which has extremely low pretest performance and indicates near complete mastery at posttest. This result is in complete agreement with the earlier conceptualization of the appropriate use of this methodology.

Furthermore, the index is dependent, although to a lesser degree, upon score variability. It should be noted that this relationship is negative, i.e., higher score variability reduces the sensitivity. This is a direct result of the precision with which the means are estimated. The more precisely the means are estimated, the more sensitive one would expect the measure to be.

No correlation was found between the sensitivity of the test and the degree of correlation between observations on each of the two occasions. This agrees with the earlier conceptualization of individual differences as an irrelevant dimension in such studies.

Finally, the ratio of the standard deviations shows no linear relationship with test sensitivity. With regard to this last observation, one might not expect the relationship to be linear. Indeed, if homogeneity of variance is important, one would expect that, other things being equal, test sensitivity would be highest when this ratio approaches 1.0 and lower as the two variances become more and more discrepant. In order to investigate this relationship further, and to more fully determine the actual shape of the previously determined relationships, plots of each of the variables with the index were obtained.

Figure 1 indicates the relationship between the variability ratio and the sensitivity. The relationship between sensitivity and the difference between pre- and posttest performance is shown in Figure 2. The graphs of the relationships between the remaining variables and the sensitivity value appear in Figures 3-7.

An examination of Figure 1 indicates that the pattern of responses shows no relationship between test sensitivity and the variability ratio. Thus, heterogeneity of variance does not appear to be seriously damaging, although heterogeneity may cause a somewhat inflated estimate of the error variance.

Results from the analysis of the relationship of the index to the two standard deviations and their ratio indicate that while sensitivity is adversely affected by overlapping pre- and posttest distributions, it is not affected by heterogeneity of variance.

From a further perusal of Figure 2 it becomes obvious that test sensitivity is determined by the prepost difference to a large extent and that this relationship is slightly curvilinear.

Figure 1

Relationship Between Sensitivity and Variability Ratio

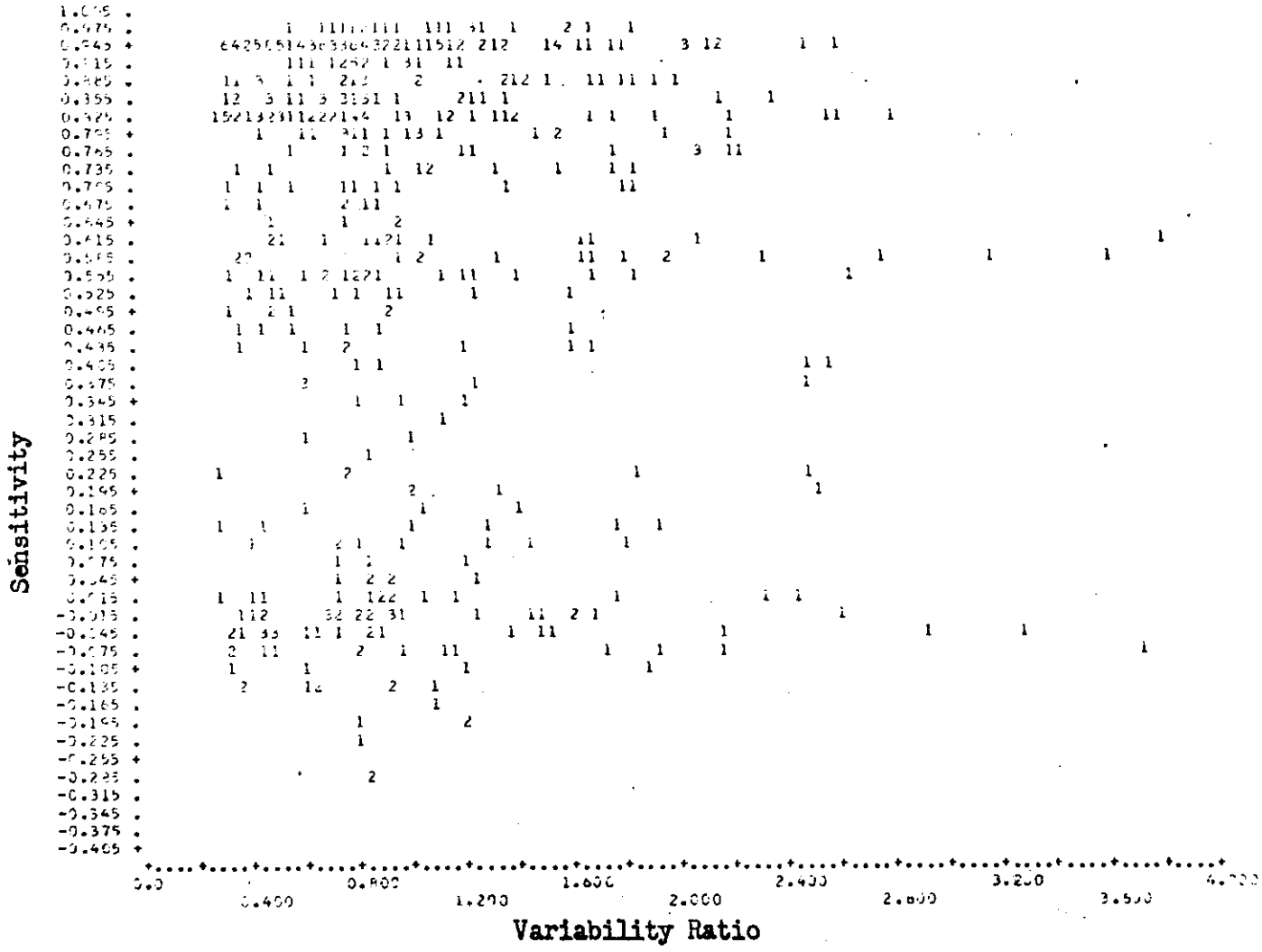


Figure 2

Relationship Between Sensitivity and Pre-post Difference

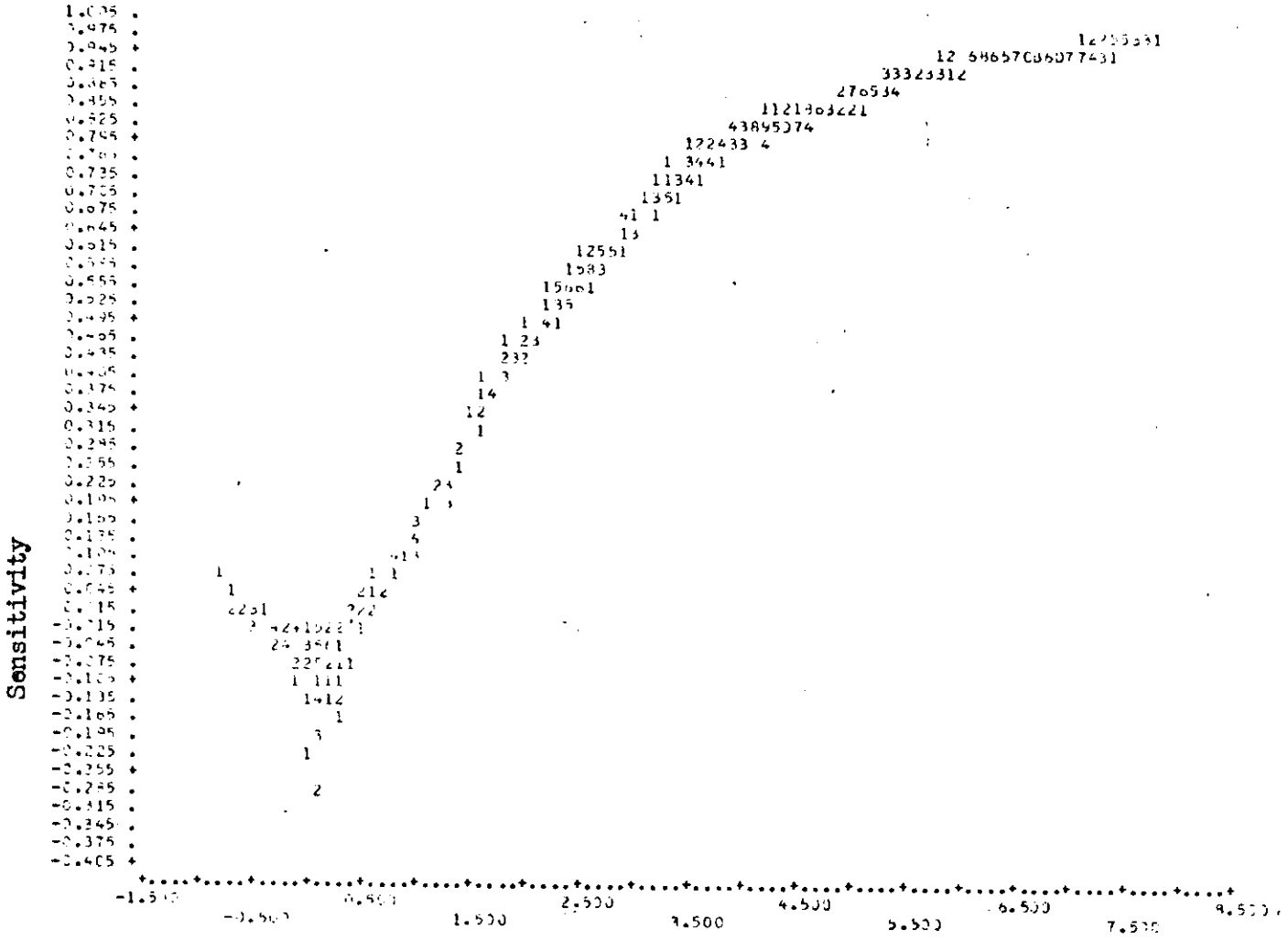
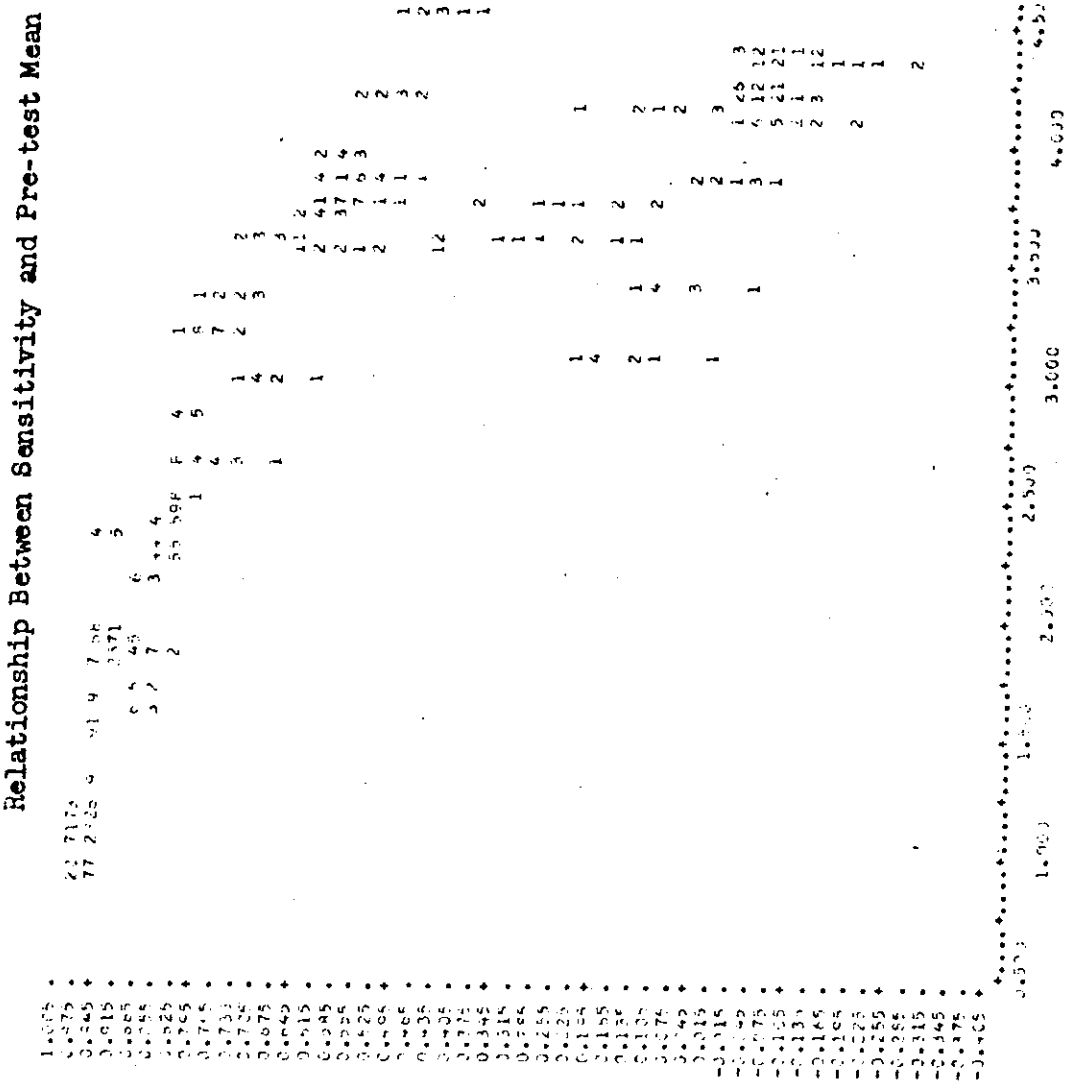


Figure 3

Relationship Between Sensitivity and Pre-test Mean



Pre-test Mean

Figure 4

Relationship Between Sensitivity and Post-test Mean

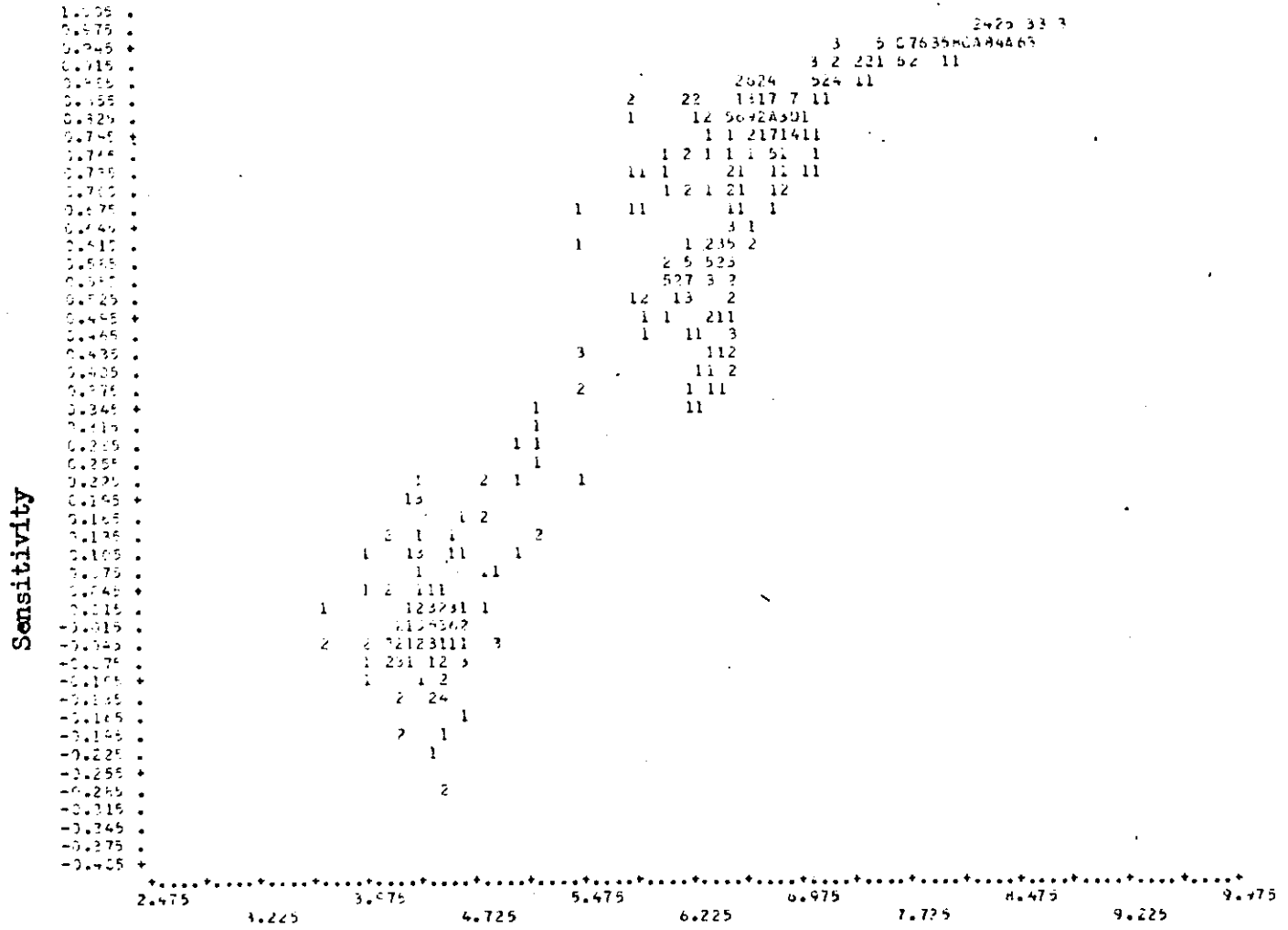


Figure 5

Relationship Between Sensitivity and Pre-test Standard Deviation

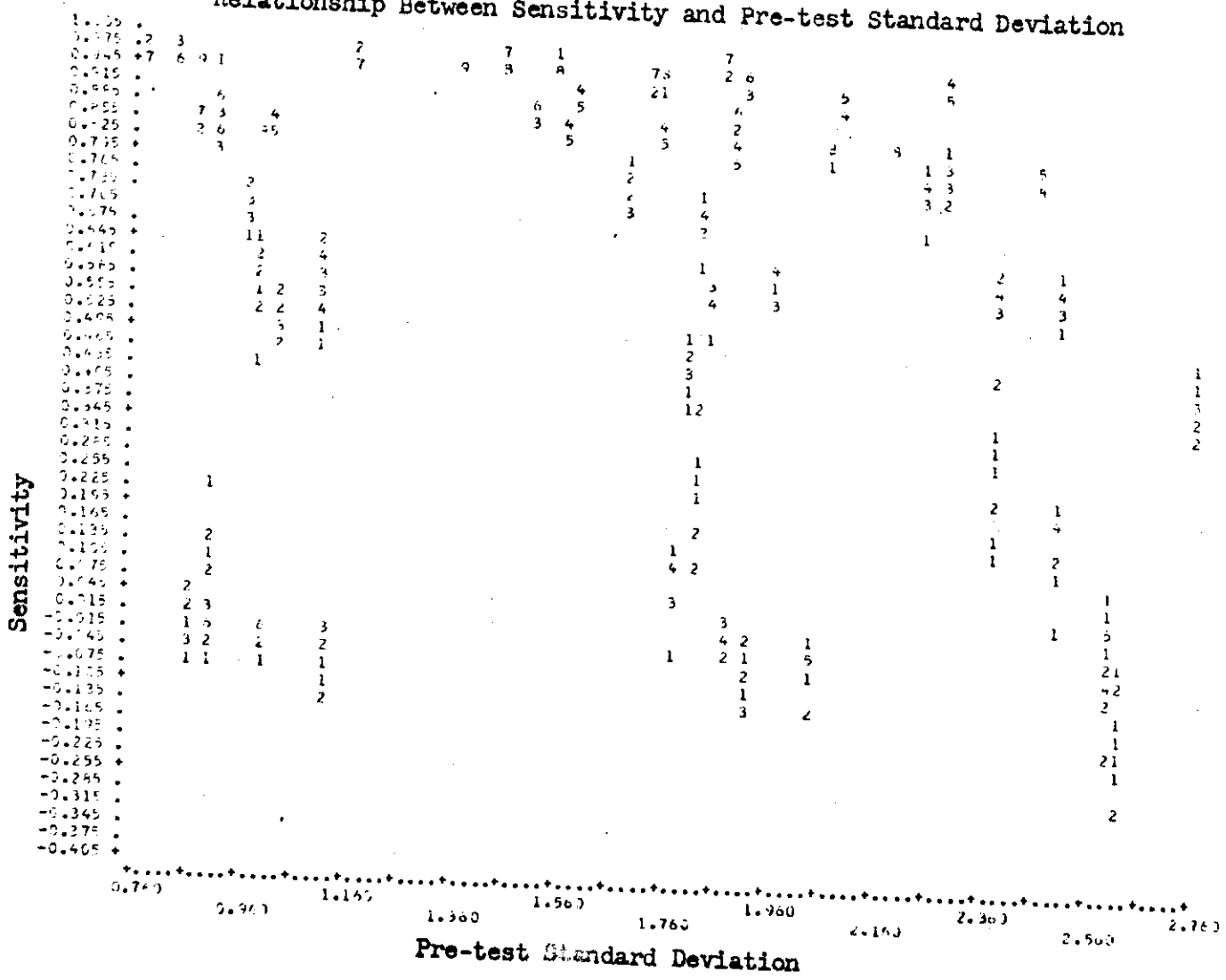


Figure 6

Relationship Between Sensitivity and Post-test Standard Deviation

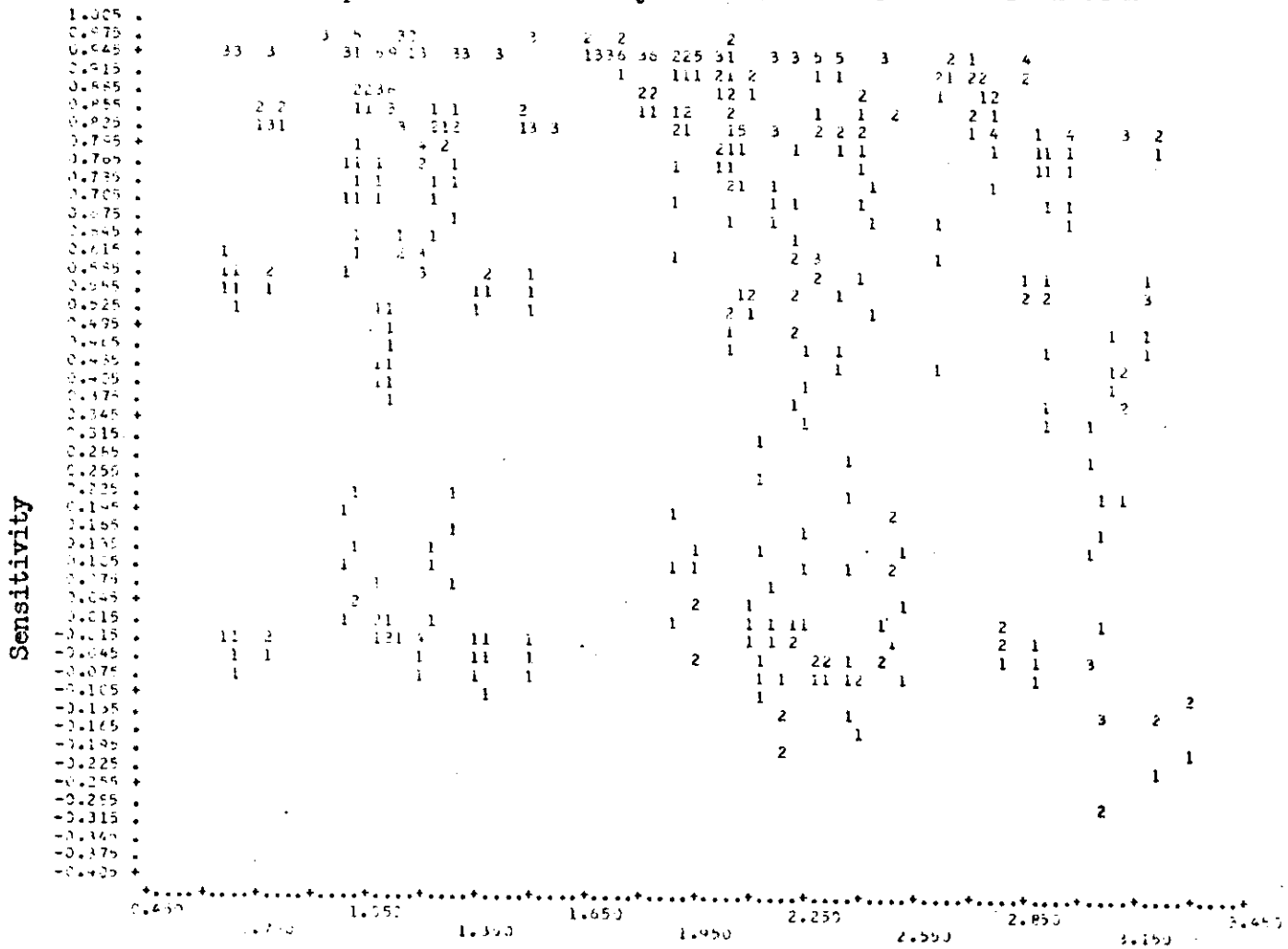
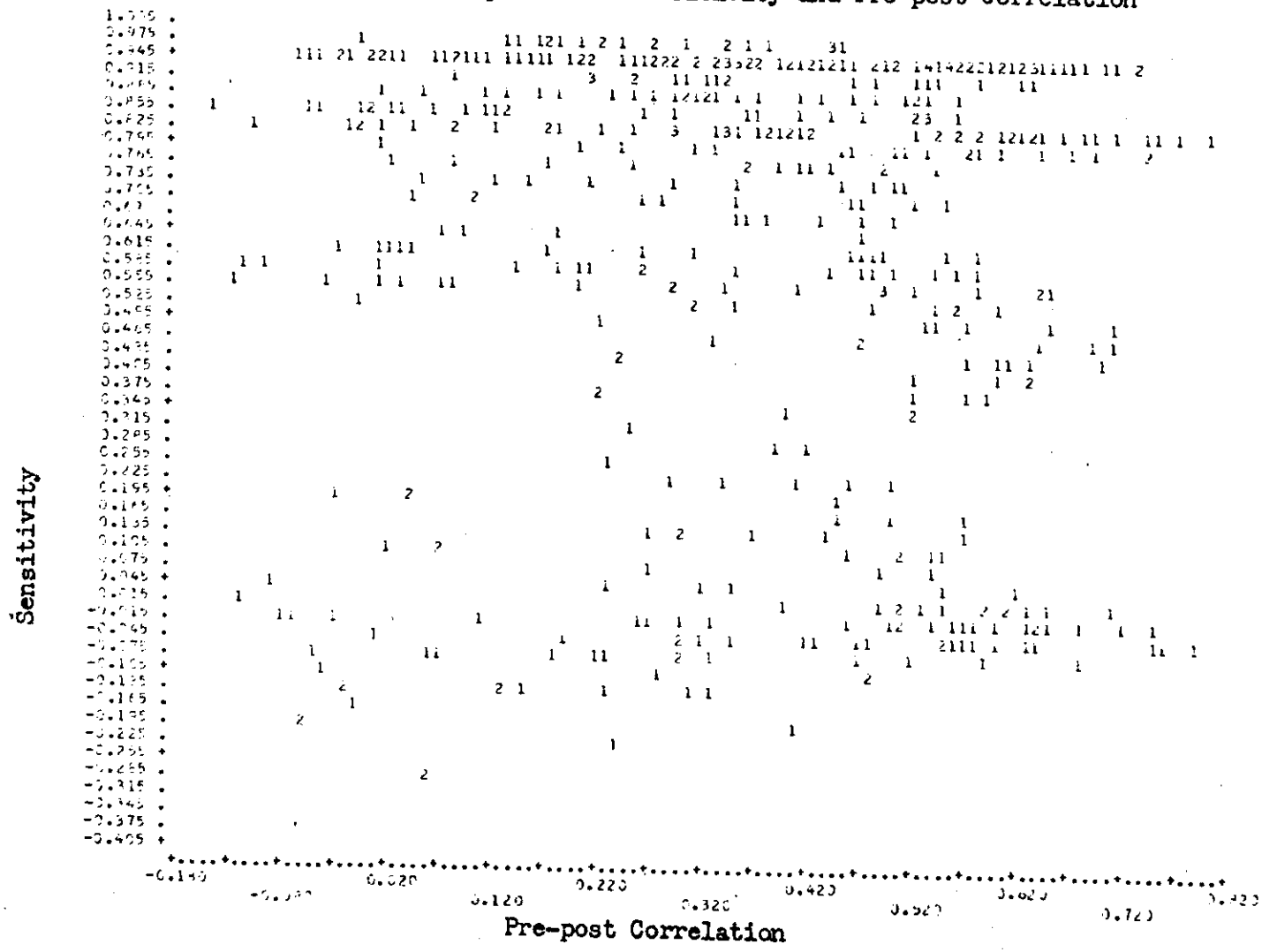


Figure 7

Relationship Between Sensitivity and Pre-post Correlation



CHAPTER VIII

METHODS AND RESULTS: EMPIRICAL STUDIES

In order to fully investigate the characteristics of the index, empirical data from a variety of sources were gathered and analyzed. It was felt that by using differing subject matters and instructional situations one could investigate not only the characteristics of the index but the generalizability of this technique as well. In the sections which follow the data from three independent sources will be described and analyzed. The alternative model which accounts for differences in objectives and methods of item selection will be introduced in conjunction with these sources.

Source 1: A Graduate Statistics Course

The first source of data comes from a test designed to measure performance on the stated objectives of a statistics course given to graduate students in educational research.[2] These data are included because they were derived from a conscious attempt to develop an adequate criterion-referenced measure.

A set of ten behavioral objectives on the general topics of probability, central tendency, and variability was written for the course and reviewed by faculty members. Six items were generated for each of the objectives and were pretested. From the pretest results a 20-item test form (designated Form A) was generated by choosing the 2 items from each of the 10 objectives with the largest value of the Cox and Vargas difference index. Although the original study was far more complex, attention here will be given only to the results of administering

2. This data supplied by Dr. Stephen Ivens, College Entrance Examination Board.

this measure to the 17 subjects used in the final phase of the study.

Table 8 indicates the effect of administering this measure before and after the instructional unit. Here, consistent with the test author's endeavor to develop a sensitive instrument to reflect the content of the instructional unit, one sees a substantial increase in performance from the first to second occasion.

TABLE 8

Summary Data for Statistics Test

	Pretest	Posttest
Mean	5.88	12.59
Std. Dev.	2.76	2.79

One's expectation is that the care involved in constructing this measure should have resulted in a measure with a substantial value of the sensitivity index.

In order to estimate the sensitivity of the measure the test data were first analyzed using the analysis of variance model associated with the first model for test sensitivity (Equations 8, 9). The analysis of the variance table below (Table 9) shows the results of this analysis. It should be noted that this initial analysis ignores differences between objectives and treats all items as replicate measures of the same broad content.

TABLE 9

Analysis of Variance of Statistics Test

Source	Degrees of Freedom	Mean Square
Occasions	1	19.112
Subjects	16	.428
Occasions X Subjects	16	.340
Error	646	.213

Using Equation 14 the sensitivity of the test was estimated as .838. This value would seem to be in agreement with the test developer's conception of the measure as a good example of a properly functioning criterion-referenced measure.

These data, because of the carefully planned construction to represent each objective, also allow for demonstration of the third model for estimating test sensitivity (see Equation 20). When the model which accounts for variability due to differences in objectives is applied to the data the result is a decrease in the estimate of the error variability. This indicates that variability originally considered to be due to error is actually due to differences in the objectives. The revised estimate of the test sensitivity now becomes .842.

Finally, and somewhat tangentially, the data can be used to demonstrate the effects of the correction for guessing. An inspection of Table 8 indicates that pretest performance is very near that expected if subjects only guessed on the items. Furthermore, it seems reasonable that in a specialized subject matter, such as statistics, that subjects

would bring very little content knowledge with them to the testing situation. Thus one might expect considerable guessing in this situation. If the correction for guessing is applied to the number of items attemted by each subject at pretest and the resulting adjusted scores are used in the analysis of variance of the first model, the adjusted estimate of the test sensitivity is .910.

Source 2: The Los Angeles Mathematics Project

Methods

The second source of empirical data comes from part of the data collected in conjunction with the evaluation of experimental mathematics programs for junior high school students in the Los Angeles City Schools.[3] While these data were not derived from a criterion-referenced measure, they are included here because they provide a large data base useful for the demonstration of item selection techniques and the effect of these techniques on the resulting test sensitivity.

Although data from a variety of sources were used in the original evaluation only the responses to the Diagnostic Test constructed by the evaluation staff are of interest here. Construction of this instrument was based on an attempt to measure the general instructional goals of the experimental program. From these goals eleven content areas were identified and several items were generated within each area. After review by the evaluation staff, the retained items were used to generate a parallel form for each item. Two forms of the test were developed

[3]For a complete description of this research see C. Wayne Gordon, Evaluation Report on the Los Angeles City Schools SB 28 Demonstration Program in Mathematics, Parts 1-3, (mimeograph) Center for the Study of Evaluation, University of California, Los Angeles.

by randomly assigning the members of each pair of items to Form A or Form B of the test.

Both forms of the test (40 items each) were administered at the beginning and end of the school year under study. Thus, the data base for this section consists of the responses of the 329 seventh-grade students for whom complete data were available, to 80 items given before and after a year of instruction.

Although instructional goals were used as the basis for the design of the measures, these data cannot be considered to be representative of adequate criterion-referenced measures for a variety of reasons. First, the goals specified were quite general, often vague, and subsumed several instructional units. Second, they represent the goals of three separate experimental programs which were combined to provide a general measure useful for comparisons in the original evaluation study. Finally, the data used here are the results of pooling six sub-populations (experimental and control groups in each of three schools). Such pooling obscures true instructional effects on those items which are appropriate for the particular sub-population. For the purposes of the present research the original use of the data can be ignored in order to investigate item selection techniques.

For present purposes subjects were split into developmental (D) and cross-validation (C) groups. Item selection strategies were then used to develop measures from the responses of the developmental group which were then applied to the cross-validational group. In order to represent the original sub-populations, stratified random samples were used to obtain the two groups ($N_D = 165$, $N_C = 164$).

Various strategies were used to develop measures with different properties. Each strategy began with an identification of the best

items. While a few alternative methods were fruitlessly attempted, the Cox and Vargas difference index was finally used to identify these items. Use of this method is consistent with a maximization of the between occasions variance and will therefore lead to maximum sensitivity. Once this realization had been achieved, it was decided to examine the effect of using this technique under various constraints.

Three approaches were used to reflect varying numbers and types of constraints. The first constraint condition was to develop a test which had a maximum sensitivity value but which represented each of the 40 original item contents. The second constraining condition was to develop a test with the maximum sensitivity value which equally represented the 11 broader content areas. Finally an attempt was made to develop a test which was maximally sensitive with no constraints. The three approaches allow one to note changes in the nature of the derived measures as a result of applying various types of constraints.

The methodology suggested by Popham (1970) for identifying atypical items was also applied in an attempt to develop a highly sensitive test by the strategy of deleting poor items as opposed to selecting good items as described above.

Results

As a first step the proportions of students passing the items on each of the two occasions for the 165 students in the developmental group were computed. These proportions are shown in Table 10. This table is organized by content areas and lists the parallel items next to one another.

Two results from these data are especially striking. One notices a substantial number of high initial means and small item differences.

The first result would indicate that the students are bringing knowledge of some item contents with them to the instructional setting. One's expectation in this case is that the items with initially low means would then show the largest increases since, clearly, these are the contents in which students are deficient. The second result, however, indicates that few items show substantial differences between occasions. Data from the evaluation of the previous school year indicate that teachers were giving instruction on skills already available to the students (Skager, 1969). One might expect the same phenomenon to be present in these data since the same teachers and programs were involved. That the items with initially higher means did not show larger gains can be attributed to such factors as a ceiling effect for the item or a lack of motivation on the part of the students to relearn previously presented material.

In all of the analyses presented in this section the first model for estimating sensitivity (Equations 8 and 13) was used in order to make comparisons in the resulting sensitivity values possible. The analysis began by computing the sensitivity for the original forms of the test and for the combined test. These values are given in Table 11. In order to develop a test by the first strategy the items showing the greatest Cox and Vargas difference value of each of the parallel pairs of items were included in the derived measure. The items selected are underlined in Table 10. This measure, designated S1, was then scored for the developmental group. The results of applying the test developed in this manner to the D and C groups are also shown in Table 11.

It should be noted that the measure developed in this manner includes three items that indicate no positive gain in the D group. This is due to the fact that neither item in the parallel pair demonstrated

TABLE 10

Item Proportions - Developmental Group

Content	Form A Item	Pre	Post	Diff	Form B Item	Pre	Post	Diff
Integers	2	.503	.624	.121[a]	22	.642	.715	.073
	5	.442	.612	.170	5	.461	.521	.060
	20	.606	.788	.182*[b]	7	.654	.733	.079
	37	.285	.442	.157	1	.352	.533	.181*
Rational numbers	27	.588	.697	.109	39	.497	.733	.136
	19	.188	.315	.127	29	.630	.709	.079
	39	.254	.297	.043	25	.242	.358	.116
	21	.164	.152	-.012	33	.182	.218	.036
	30	.721	.685	-.036	16	.836	.842	.006
	25	.521	.648	.127	20	.321	.509	.188*
	24	.382	.588	.206*	12	.503	.461	-.042
Measurement	22	.224	.254	.030	10	.721	.800	.079
	29	.412	.509	.097*	19	.333	.364	.031
	8	.539	.521	-.018	21	.558	.546	-.012
	18	.200	.279	.079	6	.254	.327	.073
	13	.309	.582	.273*	3	.515	.685	.170
Algebra	40	.194	.346	.152*	38	.461	.521	.060
	35	.582	.709	.127	32	.418	.588	.170*
Geometry	1	.327	.515	.188*	15	.206	.236	.030
	33	.612	.636	.024	27	.327	.364	.037
	17	.685	.642	-.043	40	.442	.570	.128*
	32	.054	.158	.104	28	.085	.170	.085
Place value	3	.400	.558	.158*	34	.382	.436	.054
	4	.400	.600	.200*	23	.327	.303	-.024
	34	.400	.418	.018	4	.709	.703	-.006
Number theory	12	.152	.212	.060	14	.394	.352	-.042
	14	.279	.582	.303*	24	.430	.673	.243
	6	.776	.843	.067	9	.733	.794	.061
	10	.067	.146	.079*	37	.394	.333	-.061
Set theory	26	.448	.382	-.066	17	.291	.370	.079*
	7	.376	.315	-.061	18	.346	.333	-.013
	28	.588	.727	.139	30	.673	.836	.163*
Field axioms	31	.254	.285	.031	35	.115	.182	.067*
	38	.261	.248	-.013	31	.297	.297	.000
	9	.624	.648	.024	26	.467	.582	.115*

TABLE 10 (cont.)

Item Proportions - Developmental Group

	Form A				Form B			
Content	Item	Pre	Post	Diff	Item	Pre	Post	Diff
Statistics	16	.139	.200	.061	11	.206	.261	.055
	11	.691	.800	.109*	13	.746	.806	.060
	36	.254	.388	.134*	8	.364	.485	.121
Word problems	15	.454	.503	.049*	36	.146	.103	-.043
	23	.594	.685	.091*	2	.582	.654	.072

[a]Underline indicates the item from each pair with the largest instructional gain.

[b]Asterisks indicate the two items from each content area with the largest instructional gains, excluding the parallel form of the first item selected.

TABLE 11

Test Sensitivity Values for Various Test Forms in Developmental and Cross-Validation Groups

Test form	# of items	Developmental	Cross-Validation
Form A	40	.408	.475
Form B	40	.286	.306
Combined AB	80	.519	.568
S1[a]	40	.534	.515
S1'[b]	37	.559	.548
S2[c]	22	.525	.464
S3(opt)[d]	42	.649	.634
S3'(opt)[e]	24	.591	.548

[a]best item from each item pair.

[b]best item from each item pair, positive differences only.

[c]best two items from each content area, no parallel items.

[d]items entered in order of size of difference index, optimum value.

[e]items entered in order of size of difference index, no parallel items, optimum value.

an increase in performance. Certainly items with negative or no difference cannot add to the ability of a test to discriminate between levels of competency. Therefore, these items were dropped to form test S1' which was then reanalyzed yielding the results shown in Table 11.

The second approach was to develop a test which equally represented all of the content areas. This test was constructed by selecting the two items with the highest pre-post discrimination from each of the content areas subject to the constraint that the two items could not be parallel forms of each other. The items selected are marked by an asterisk in Table 10. The test form thus constructed is designated S2. The sensitivity values computed when scored for the developmental and cross-validation groups are given in Table 11.

Finally, an attempt was made to maximize the value of the sensitivity index while disregarding content areas. First all 80 items were ranked in terms of the magnitude of the pre-post difference index. Starting with the 4 items with the highest value of this index, tests were constructed and analyzed by adding the two items with the highest difference index at each successive stage. The largest test generated in this manner contained 50 items. It was decided to stop at this point since the value of test sensitivity had reached its maximum and had begun to decline and because a test of greater than 50 items would seem too lengthy to use in a practical setting. The sensitivity values for each test thus constructed are shown in Table 12. The equivalent values for the cross-validation group are also shown here. As expected there is some shrinkage of the index when applied to this group but it is not large. Furthermore, the fluctuations observed in the value of the index near its maximum may be attributable to a slightly

different ordering of items in the cross-validation group. The optimum test lengths and their associated sensitivity values are shown in Table 11, and are designated as S3 (opt).

The methodology described above allowed parallel items to be included in the test. An alternative approach, S3' (opt), again added items two at a time but restricted this inclusion to only those items that did not have a parallel form already included in the test. The values of the sensitivity index for tests of various lengths constructed in this way are shown in columns 4 and 5 of Table 12. The optimum values of this modified strategy are designated S3' (opt) in Table 11. It should be noted that there is a decrease in the value of test sensitivity using this strategy. This is not surprising since the exclusion of parallel item forms does not allow those content areas where large gains were made to be weighted more heavily.

The application of Popham's method for identifying atypical items was applied to the items in each of the 11 content areas. In this methodology the frequency of occurrence of the four possible outcomes for an item given on two occasions is tabulated across subjects. Then within a group of items from the same objective the median frequencies for each category are computed. These medians thus represent the typical item from that objective. Chi-square values are then computed for each item. Thus items which differ greatly from the pattern of responses for a typical item will have large chi-square values.

In the absence of instructional objectives, the 11 content areas were used to define sets of items. Within each set, the median frequencies and chi-square values were computed. The frequencies of each of the possible response patterns and the chi-square values for each

TABLE 12

Test Sensitivity as a Function of the Number of Items[a] Included in the Test.

# Of Items	Parallel items		No parallel items	
	Developmental	Cross-valid.	Developmental	Cross-valid.
4	.376	.403	.362	.271
6	.435	.406	.424	.350
8	.481	.414	.461	.367
10	.512	.465	.498	.450
12	.539	.515	.525	.485
14	.563	.543	.565	.501
16	.582	.550	.571	.501
18	.610	.582	.581	.522
20	.615	.581	.586	.533
22	.626	.595	.588	.540
24	.633	.599	.591*	.548*
26	.638	.613	.589	.540
28	.644	.623	.589	.540
30	.646	.626	.583	.530
32	.645	.632	.568	.533
34	.646	.633		
36	.646	.620		
38	.646	.626		
40	.647	.623		
42	.649*	.634*		
44	.645	.628		
46	.643	.634		
48	.634	.633		
50	.633	.632		

[a]Items entered in order of size of the pre-post difference index.

*Indicates maximum value

item, grouped by content areas, are shown in Table 13. It is obvious that the items within each of the content areas are not very similar. This is not surprising when one remembers that these items were not generated as replicate measures of stated instructional objectives. In the next section an example will be presented where this methodology is applied to items generated from objectives and expected to be highly similar. In later discussion the differences in these two applications will be compared.

The failure of this technique to identify a relatively small number of atypical items resulted in a decision to abandon an attempt to develop a more sensitive test by identifying poor items for these data.

The application of various strategies for constructing a measure has led to tests with varying sensitivities. It has been shown that selecting items from the same pool, but under varying restrictions, leads to somewhat different measures. The most sensitive measure, S3(opt), was the one that placed virtually no restrictions on the kinds of items included. This strategy allowed one to capitalize on those content areas where there had been effective instruction. Restricting the measure to only one item from such item contents (S3'(opt)) lowered this value somewhat. Picking the best item from each pair of items (S1) lowered the sensitivity value even further because it forced the inclusion of relatively poor items. By comparison, deletion of the three negatively discriminating items (S1') helps somewhat, but leaves many poor items in the test. The lowest sensitivity for the derived measures comes from the measure with the greatest restrictions (S2). Here, each content area had to be equally represented and no pairs of parallel items could be included. It is interesting to compare the

TABLE 13

Response Frequencies and Chi-Square Values for Mathematics Test Items

Content	Item	00[a]	01[b]	10[c]	11[d]	Chi-Square
Integers	A2	81	89	45	114	0.48
	A5	79	108	41	101	7.02
	A20	36	88	38	167	51.57
	A37	154	77	27	71	89.37
	B22	47	57	38	187	73.63
	B5	84	83	52	110	2.96
	B7	35	60	42	192	90.06
	B1	113	97	45	74	28.34
Rational numbers	A27	64	76	32	157	47.32
	A19	196	72	29	32	154.88
	A39	169	74	47	39	94.39
	A21	240	40	34	15	307.78
	A30	35	61	71	162	105.84
	A25	70	87	43	139	19.09
	A24	103	90	37	99	4.61
	B39	77	84	19	149	42.47
	B29	46	68	35	180	94.92
	B25	146	98	55	30	88.35
	B33	213	57	46	13	224.89
	B16	13	47	46	223	240.67
	B20	135	81	32	81	22.15
	B12	87	73	72	97	26.28
Measurement	A22	208	52	27	42	140.33
	A29	98	89	63	79	17.80
	A8	103	57	67	102	21.94
	A18	187	70	46	26	109.63
	A13	96	128	28	77	65.08
	B10	23	63	36	207	274.34
	B19	161	58	47	63	40.53
	B21	86	59	66	118	35.55
	B6	174	68	31	56	63.40
	B3	64	101	40	124	59.83
Algebra	A40	196	59	20	54	33.59
	A35	51	67	40	171	172.53
	B28	106	70	62	91	45.30
	B32	83	104	41	101	60.72

TABLE 13 (cont.)

Response Frequencies and Chi-Square Values for Mathematics Test Items

Content	Item	00	01	10	11	Chi-Square
Geometry	A1	116	106	31	76	45.07
	A33	68	59	52	150	156.96
	A17	47	57	79	146	209.62
	A32	268	40	11	10	218.18
	B15	210	45	36	38	63.43
	B27	142	75	63	49	26.54
	B40	110	86	39	94	32.25
	B28	261	39	12	17	191.75
Place value	A3	106	79	30	114	24.02
	A4	83	92	57	97	19.38
	A34	124	65	62	78	1.89
	B34	124	72	54	79	1.68
	B23	174	47	59	49	50.62
	B4	34	56	69	170	140.57
Number theory	A12	213	65	33	18	117.70
	A14	108	130	25	66	83.49
	A6	25	49	24	231	487.21
	A10	252	51	21	5	222.54
	B14	123	66	80	60	76.30
	B29	64	122	38	105	111.41
	B9	32	55	30	212	383.82
	B37	152	47	64	66	47.88
Set theory	A26	130	54	77	68	8.84
	A7	138	71	76	44	6.39
	A28	49	91	34	155	232.08
	B17	151	71	58	49	4.09
	B18	161	57	68	43	12.70
	B30	23	84	28	194	427.62
Field axioms	A31	183	71	54	21	6.77
	A38	178	62	52	37	1.65
	A9	57	69	48	155	495.19
	B35	245	43	33	8	70.72
	B31	160	75	62	32	3.89
	B26	96	76	43	114	216.22

TABLE 13 (cont.)

Response Frequencies and Chi-Square Values for Mathematics Test Items

Content	Item	00	01	10	11	Chi-Square
Statistics	A16	237	49	19	24	107.91
	A11	23	79	38	189	416.16
	A36	163	77	47	42	15.09
	B11	218	56	25	30	68.07
	B13	30	55	32	212	524.73
	B8	106	101	53	69	36.47
Word problems	A15	81	97	62	89	38.06
	A23	71	71	32	155	13.32
	B36	268	26	30	5	620.13
	B2	68	69	38	154	13.18

[a]fail-fail [b]fail-pass [c]pass-fail [d]pass-pass

sensitivity of this 22-item test (.525) with the value obtained for S3 when the best 22 items have been entered (.626).

The main implication of these findings is that the nature of the test changes as a function of the external restrictions placed upon its form. Clearly, the more restrictions one must place on the measure in terms of such concerns as representation of differing contents or number of items, the less opportunity one has to capitalize on those contents where instruction produced performance increases. This conclusion must be tempered somewhat with a consideration of the special nature of these data. Ordinarily one might expect larger and more consistent changes for items written for specific instructional objectives. In that case item selection would be a matter of the selection of the best items from a pool of items which all show instructional increases. Here, one had to pick the few items which demonstrated an increase. Perhaps with more adequate items, the differences in the sensitivities of the derived measures would not have been so dramatic.

Source 3: The Denver Data

Methods

In addition to the sources previously presented it was possible to obtain data from a third source which represents somewhat of a compromise between the approaches previously presented. In the first source, although somewhat limited in sample size, data were obtained on a carefully constructed criterion-referenced measure. In the second source, data were related to an attempt to measure only very general instructional goals and were derived from measures constructed as a norm-referenced test. In this, the third source, it was possible to gather

data from tests designed to measure the stated objectives of an instructional unit but with items which had not been pretested or refined in any way.

The data presented here came from teacher-made tests given before and after appropriate instructional units at the elementary school level. In the school[4] under study a team teaching approach is used, making it possible to get fairly good-sized samples of students all of whom had been exposed to the same instructional unit.

While the measures to be analyzed here represent two content areas both were constructed in a similar fashion. First the instructional teams specified the objectives of each instructional unit. Items measuring each objective were then generated and pre- and posttest forms constructed. In this way each instructional unit had both specified objectives and items designed to measure those objectives.

The two instructional units selected for study here involved phonics and geometric concepts. Data were derived from test forms obtained from the teachers' files. These units were selected because fairly complete data were available (for many units the graded posttests had been sent home with the children). Additionally, the subject matter and grade level of these units adds some variety to the data reported in previous sections.

In both of these measures the same general format prevails. Each represents what the teachers indicated as two complexity levels (specified C and D) of the relevant content. The objectives are coded to represent the content area and level. Thus the objective

C - G - 33c: Recognizes the point of intersection of two lines.

represents objective c of level C of content area G-33.

---

[4]Eastridge Elementary School, Cherry Creek School District, Denver, Colorado.

The test forms for the two contents differed to some degree in their construction. For the phonics test, parallel items were generated for each measure. For the geometric concepts unit, some of the items were identical while others differed only in labelling.

The method of analysis to be used here begins by presenting the item proportions for each measure on the two occasions. The sensitivity of each measure was then computed. The two levels in each content area were first treated as subtests (and sensitivity values computed for each) and then combined to form a composite measure (for which a separate sensitivity value was computed).

The phonics measure represents an instance where each of several objectives is measured by several items. Therefore the model for accounting for objectives will be used on this measure and compared with the results obtained if one ignores the objectives.

Popham's method for identifying atypical items will be used in conjunction with the geometric concepts measure. Since it is felt that these data represent a situation more appropriate for the use of this methodology, they will provide a basis for comparison with the results obtained from the previous data.

Furthermore, the phonics test presents a new situation in that each objective has an associated item format with a large number of elements in the appropriate replacement set. Since the same items were not given on both occasions but parallel items were generated from the replacement sets for the item formats, the difference values computed for each item under each objective are a function of the items picked to form each pair. For this reason only the average gain for each item format (i.e., objective) will be computed.

Results

The proportion of students passing each item on each of the two occasions was computed for each measure. These are listed in Tables 14 and 15. For the phonics test the average levels of performance on each objective within each occasion were also computed and are listed along with the average increase in performance. Since virtually the same items were used on both occasions for the geometric concepts measure, the differences between levels of performance for each item are listed for this test.

The computed sensitivities for each measure are listed in Table 16. For the phonics measures four values appear. The test designated C7 represents all phonics items at the C level. Test C7D7 represents the items at both the C and D levels. It should be noted that there are fewer subjects listed for the combined measure. This is due to the fact that the two measures were given on separate forms and the D7 post-test forms were not available for some students (they were sent home with students).

The tests designated as C7-OB and C7D7-OB represent the same measures and subjects as above but reflect the use of the model which accounts for variability due to objectives. These results indicate that the sensitivity is increased when one accounts for variability due to objectives. An inspection of the values in Table 13 verifies that differences between objectives exist. In particular one notes a considerably lower level of performance on the first objective than on the other five. Even among the remaining five there is considerable variability as indicated by the objective means, although this is not so dramatic as it is for the first objective.

TABLE 14

Item Proportions for Phonics Items

Objective	Item	Pre	Post	Diff.
C7a	1	.064	.819	
	2	.064	.532	
	3	.053	.511	
	4	.915	.936	
	5	.053	.543	
	Ave.	.230	.668	.438
C7b	1	.894	.947	
	2	.287	.872	
	3	.277	.894	
	4	.479	.925	
	5	.839	.505	
	Ave.	.553	.826	.273
C7c	1	.787	.617	
	2	.500	.957	
	3	.723	.883	
	4	.766	.819	
	5	.436	.883	
	Ave.	.643	.832	.189
C7d	1	.936	.968	
	2	.681	.862	
	3	.883	.649	
	4	.745	.979	
	5	.638	.851	
	Ave.	.777	.862	.085
D7a	1	.354	.973	
	2	.378	.973	
	3	.744	.703	
	4	.427	.568	
	5	.317	.460	
	Ave.	.600	.735	.135
D7e	1	.646	.865	
	2	.634	.595	
	3	.537	.649	
	4	.500	.784	
	5	.439	.730	
	Ave.	.600	.724	.124

TABLE 15

Item Proportions for Geometry Items

Objective[a]	Item	Pre	Post	Diff.
CG33a	1	.973	.973	.000
	2	.947	.987	.040
	3	.960	1.000	.040
	4	.027	.973	.946
CG33c	1	.853	1.000	.147
CG33f	1	.000	.987	.987
	2	.000	.987	.987
	3	.000	.973	.973
CG33g	1	.013	.960	.947
	2	.027	.867	.840
	3	.067	.893	.826
DG33b	1	.027	.860	.833
	2	.000	.860	.860
	3	.000	.880	.880
	4	.000	.900	.900
DG33f	1	.000	.660	.660
	2	.000	.760	.760
DG33j	1	.053	.940	.887
	2	.027	.840	.813
	3	.013	.920	.907
	4	.000	.880	.880
	5	.013	.960	.947

[a]Level C proportions are based on 75 subjects, level D on 50 subjects.

TABLE 16

Sensitivity Values for Objective-Based Measures

Content	Measure	# of items	# of students	Sensitivity value
Phonics	C7	20	94	.760
	C7D7	30	31	.683
	C7-OB	20	94	.798
	C7D7-OB	30	31	.739
Geometry	CG33	11	75	.937
	DG33	11	50	.983
	CGDG33	22	50	.980

(OB indicates model for objectives has been used)

The results of estimating the sensitivity for the geometry unit are also included in Table 16. Here separate estimates are given for each of the items at the two levels as well as the estimate for the aggregate. Here one notes the disparity between the estimates for the C level items and the D level items and between the C level and the aggregate of C and D items. But one need only look at the item proportions of Table 15 to see why this is so. With the exception of the first three items, and the fifth, all items in this test represent contents about which the subjects demonstrated practically no knowledge prior to instruction and almost complete mastery after instruction. The four items that break from this pattern indicate near complete mastery prior to instruction. The effects of such a disparity in the type of items in the test are first a lowering of the between occasions variance and secondly an inflation of the error variance. The net result is a reduction in sensitivity. When the two levels are combined, these four items constitute a minor number of the items and their negative effect results in only a slight reduction in the overall sensitivity. It should be kept in mind that this reduction, although small, comes with a doubling of the test length. Ordinarily one would expect that the increase in test length with items similar to those already included would be attended by a decrease in error variance and hence an increase in sensitivity.

The present example also allows for a rather dramatic demonstration of Popham's methodology for identifying atypical items. If all 22 of the items are considered to be measures of the same general objective then one could use the methodology previously described to obtain chi-square values for the extent to which the response patterns

for each item diverge from the median or typical values. The frequencies of each response pattern for each item, as well as the chi-square values, are listed in Table 17.

Clearly the first three items, as well as the fifth, are quite different from the rest of the items which comprise the test. In this case it is easy to tell why this is so. All of the remaining items show predominantly a pattern of fail at pretest and pass at posttest. This is the type of item that ideally should be included in a criterion-referenced measure. The four items with the extraordinarily high chi-square values are alike in that they all represent subject matter which the children already knew prior to instruction. The implications of certain aspects of these outcomes will be discussed later in more detail.

Items 16 and 17 differ from the other items in that, although they show improvement between the two occasions, learning was not as complete as with the other items. This may imply that the concept these items measure was not taught as well as other concepts in the instructional unit. The last item (Item 22) is different from the other items in that a larger number of students already knew the behavior measured by this item prior to instruction.

Also of interest in Table 17 are the entries in column "10". The entries of this column can be considered as observable errors. If one assumes that on each occasion a subject is either able or unable to solve a particular problem, and that only positive changes in ability occur, then any entry in this column must come from a response error on one of the two occasions. There is only one such error in these data. This may be compared with the results of the mathematics test in Table 13. One of the reasons for the differences in the results of

TABLE 17

Response Frequencies and Chi-Square Values for Geometric Concepts Items

Level	Item	00	01	10	11	Chi-Square
CG33[a]	1	0	0	0	50	2447.12
	2	0	1	1	48	2301.14
	3	0	0	0	50	2447.12
	4	1	47	0	2	2.70
	5	0	5	0	45	1972.70
	6	0	50	0	0	5.12
	7	0	50	0	0	5.12
	8	0	50	0	0	5.12
	9	1	48	0	1	1.91
	10	4	45	0	1	0.43
	11	3	43	0	4	9.02
DG33	12	7	41	0	2	6.45
	13	7	43	0	0	6.35
	14	6	44	0	0	4.03
	15	5	45	0	0	2.43
	16	17	33	0	0	68.72
	17	12	38	0	0	28.62
	18	3	43	0	4	9.02
	19	8	41	0	1	8.45
	20	4	45	0	1	0.43
	21	6	44	0	0	4.03
	22	2	39	0	9	64.74

[a]Only subjects for whom both subtests were available are included.

the two tables is that in this test responses must be produced, not just selected. The math test was in a multiple-choice format. If only chance is operating on each occasion for a four-part multiple-choice item the expectation of a correct response at pretest is one in four. The expectation of the joint occurrence is three in sixteen. This would lead to an expected frequency of about sixty-two persons on the math test. Most items do not exceed this value.

The implications with regard to item selection and test usage of various patterns in these response outcome tables will be discussed in the next chapter.

CHAPTER IX

DISCUSSION

The previous sections have described several sources of data, both simulated and empirical, which have been used to study the sensitivity concept under a variety of conditions. The separate results of these several studies will be used to formulate some general conclusions regarding the sensitivity concept. Finally, some suggestions for further research will be made.

From the variety of sources presented here the most apparent result is that the sensitivity increases as pretest and posttest distributions become less and less overlapping. This result is consistent with the initial conceptualization of the sensitivity as a measure of the test's ability to discriminate competency levels. Certainly as the performances of the competency groups become more distinct, the measure is better able to classify the subjects' performance on the test into one of these two groups. While the classification problem has not been approached in this paper, it should be clear that if a student scores at or above the high competency group mean when given the measure prior to instruction, he will probably benefit very little from that instruction. (Suggestions for further research on the classification problem will be given later.)

In terms of selecting items, two points are especially important. First, it has been noted that selection of items by the value of the Cox and Vargas difference index is most consistent with a maximization of the sensitivity. Secondly, Popham's method for identifying atypical items has been shown to be useful when a certain amount of item homogeneity is present. In the example of the geometric concepts test it worked well,

while with the mathematics achievement test almost every item seemed atypical. The reason is that this approach depends upon considerable homogeneity in the set of items. This homogeneity is necessary to define the "typical" item. Unfortunately, homogeneous items are not necessarily good items in the sense of great pre- to post-instructional change. A certain amount of judgement therefore is still required in the selection of items by Popham's method. Indeed one would rather select the two items out of ten which demonstrate sensitivity to instruction than the eight which are homogeneous because they measure a behavior irrelevant to the unit under study.

Since the approach presented here is an extension of the traditional response model, most of the same restrictions and considerations inherent in norm-referenced measurement theory still apply. Thus, no completely adequate statistical decision model for selecting items is available. One must select items not only with regard to such values as the difference index, but with regard to such considerations as test length, adequate representation of certain contents, and test format. There are, however, some guidelines which may be useful.

Probably the most useful approach to constructing an adequate criterion-referenced measure is to administer a relatively large-sized sample of parallel items to a group of subjects similar to those for which one wishes to construct the final measure. After the test has been administered on the two occasions the fourfold response outcomes should be tabled along with the difference index value. All of these values are useful in the study of items. The column indicating the frequency of pass-fail responses gives some indication of the number of response errors. If this value is excessively large for any one item, the item should be inspected for ambiguities in the alternatives.

An example of this kind of item would be one in which one of the alternatives becomes a confusing but possible answer on the basis of the information provided during the instructional unit. While the fail-pass column provides information similar to the difference index, the pass-pass and fail-fail columns give some indication of the difficulty level of the item. For example, a high number of pass-pass responses would indicate that the students already know the subject content which the item measures. This may indicate that the content is not the proper concern of the present instructional unit. It would certainly seem inefficient to provide further instruction in an area where students already demonstrate a high degree of competency.

With regard to the question of item selection, one must consider the problem of the effectiveness of instruction. For example, if an item demonstrates a relatively low difference index, how can one determine whether this is due to instruction or item inadequacy? It is with regard to this question that comparisons of the differences in the item response patterns become especially meaningful. First, it should be remembered that a low difference index will correspond to a relatively low frequency in the fail-pass column of the response matrix. Then, one must examine the remaining columns to identify the deficiency. A high frequency in the pass-pass column indicates the concept has been previously learned. Such information would most probably lead to a rejection of both hypotheses regarding item deficiency and would lead instead to a re-evaluation of the content domain. A high frequency in the pass-fail column would indicate an item deficiency as previously discussed.

It is only with a high frequency in the fail-fail column of the response matrix that one would come to suspect instructional inadequacy. Both hypotheses could lead to a large number of responses in this column.

If the students have not been taught the appropriate content then one could expect a large number of fail-fail responses. However, one also could certainly expect a large number of such responses if instruction had been adequate but the item was so poorly written as to exclude a correct response.

When an item appears with a large number of such responses, two areas need to be investigated. First, one must consider the validity of the item. One must question whether the item does, in fact, measure the objective. The relevant concepts here are content validity and/or objective-item congruence (Dahl, 1971). Secondly, once one is confident that the item is a valid measure of an objective of the instructional unit, then one can compare it with other items. If other items measuring the same objective show more desired patterns of responses, then the item must be held suspect. If all items measuring the same objective show this pattern then instruction should be held suspect.

If a model is used which accounts for several related objectives or content areas, then one can compare average performances in each objective or area to determine the relative adequacies of instruction over the different areas. This kind of information would be extremely useful as feedback to those responsible for the design and execution of the instructional unit.

Before discussing uses in more detail a point must once again be emphasized. The concept of sensitivity is never completely separable from the instructional effect. This apparent deficiency must, however, be considered a pseudo-problem. In norm-referenced reliability studies one assumes that there is some continuum along which it is important to make distinctions among individuals. Reliability is then dependent upon

variability between individuals with respect to the continuum. The concept of sensitivity is based upon an extension of the same model used in norm-referenced test theory and is therefore subject to some of the same restrictions. Here one is interested in discriminating competency levels with respect to some content area. Obviously, differences in competency levels must exist for the measure to be sensitive. The approach presented here has restricted itself to considering differences in competency levels to be a function of instruction. Such a restriction is of course not necessary if levels can be identified in the population. What is important to the notion presented here is that the purpose for which one is interested in the sensitivity of the measure is to be able to identify the differences between those needing and not needing the instructional unit in question. If a test is not sensitive to the effects of the instructional unit because that instruction was totally in effective, then it would seem foolhardy, at best, to make decisions about who should or should not be subjected to that instructional unit.

In this respect the most defensibel and obvious use of an adequate measure is to make decisions regarding the placement of children in instructional units. Assuming that a measure has been constructed to measure the stated objectives of a specified instructional unit and has been found to be highly sensitive when used in some previous test population, one could use the measure to indicate whether a child needs a particular unit. Surely, if his performance is at a level like that of the high competency group he can not be expected to gain much from the instructional unit.

A second use of the concept of sensitivity is in the development of a measure which accurately reflects a particular instructional unit. Selecting items which maximize the calculated sensitivity, as was done with the junior high mathematics data, assures a measure which is sensitive to

what is actually being taught. Using the item selection techniques presented in relation to particular objectives, as with the statistics test data, assures the proper content balance while at the same time providing an adequate measure by deleting inadequate items.

Thirdly, the sensitivity concept would seem to be useful in the selection of measures. For example, if several separate measures all purpose to be appropriate measures for a particular content unit, one could administer these measures and select the one with the highest calculated sensitivity as being the most appropriate measure of the particular instructional unit. While there may be many other uses of the methodology presented here, these three are certainly the most obvious. Each would seem to be a useful application in light of the current demands for an evaluative methodology for this class of measures.

The item selection and test usage notions can probably best be summarized by an example. For this purpose the example of the geometric concepts test might be informative. By the calculated value of the sensitivity this would seem to be a highly adequate measure. But one may question why this is so. First, the test was generated to measure specific objectives of the instructional unit. Second, and this point has implications for future test analysis, the item responses on this measure were generated responses. This should be compared with the more common multiple-choice measure where item responses are selected from among a limited number of alternatives. The production of responses fits more closely Harris' (1971) conceptualization of the nature of a true criterion-referenced measure. The important result of this type of response is a reduction in error variance. The error variance is considerably inflated when guessing from a limited number of alternatives is involved.

That this effect can be especially damaging can be inferred by a comparison of the phonics and geometric concepts measures. These two measures were constructed by the same teaching team and were administered to the same population of children. While the inference cannot be entirely valid because of the different nature of the subject matter, the lower sensitivity of the phonics measure can, in part, be accounted for by a greatly increased error variance. While several alternative hypotheses might equally well explain these results, one hypothesis which cannot be dismissed out of hand is that the increased error is due to the fact that the children had only to select from certain limited options. The guessing involved in such a process may contribute heavily to the error variability. In many cases the item had only two alternatives and one might therefore expect considerable error variability.

Finally, the geometric concepts test was an adequate measure because the major proportion of its items demonstrated the desired patterns. Those four items that were quite atypical indicated prior learning. If these items were eliminated, a nearly perfect measure would result. The test, in this case, could readily be used to decide whether or not students needed to take this unit, although one would not ordinarily expect a large number of primary level children to demonstrate prior knowledge of most geometric concepts.

As with most kinds of developmental research, a great deal more research needs to be done. First, the usefulness of the approach presented here can only be determined by its use in a wider variety of situations and conditions. While both the simulation study and the empirical data suggest that the methodology is a useful one and that the suggested index has desirable characteristics, the data have been somewhat restricted. Additional data sources may help to provide information regarding the

relative importance of the various test parameters beyond those presented here. The optimum test length under different conditions may also be of interest and could possibly be examined by additional simulation methods.

It has been suggested that one of the possible uses of a highly sensitive test is to make dicisions regarding the competency level of students. One area for further investigation could be the methods by which such decisions are made. A Bayesian approach might be a possible alternative. For example, Bayesian decision rules might be based on the relative probabilities that a subject's response pattern came from the distribution of responses for high competency subjects or from the distribution of responses for low competency subjects.

The relation of the present methodology to such concepts as the reliability of gain scores has not been investigated here since present concern has not been along the individual differences dimension. This too could be of interest since it may be that it is possible to have measures which include both concepts.

Summary

The response model presented has led to a notion of the sensitivity of a measure to the differences between competency groups. That the difference between performances of the two groups is the crucial factor in this concept has been demonstrated through the use of controlled, computer simulated data. When the subjects in the two competency groups are the same students before and after instruction, the correlation between scores on the two occasions has no effect on the sensitivity of the measure. This result would seem to indicate independence from the individual differences dimension which has been conceptualized as an irrelevant dimension in this methodology.

The application of the index to a variety of data sources led to several observations regarding item selection and test usage. Perhaps most important among these are the endorsement of the Cox and Vargas technique for item differences and the use of Popham's fourfold outcome table for use in item analysis. Additionally, the distinction between produced and selected responses suggests that when produced responses are not possible, corrections for guessing might lead to a more accurate estimate of the sensitivity of the measure.

The concept of sensitivity as presented here would seem to correspond to most test writers' notions of an adequate criterion-referenced measure. By this methodology a test that approached the ideal form for a criterion-referenced measure would become more and more sensitive. Near the extreme, the calculated sensitivity would approach its upper limit while traditional test indices, such as the reliability coefficient, would approach their lower limit or become undefined. On the basis of the theory and research presented here it is suggested that sensitivity is the appropriate concept for use in evaluating criterion-referenced measures and that the methodology presented here for estimating a sensitivity index is a useful technique in such evaluation.

REFERENCES

American Psychological Association. Technical recommendations for psychological tests and diagnostic techniques. Supplement of Psychological Bulletin, 1954, 51.

Brown, F. G. Principles of educational and psychological testing. Hinsdale, Ill.: Dryden Press, 1970.

Coulsen, J. E. & Cogswell, J. F. Effects of individualized instruction on testing. Journal of Educational Measurement, 1965, 2, 59-64.

Cox, R. C. Evaluative aspects of criterion referenced measures. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Cox, R. C. & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February, 1966.

Cronbach, L. J. Essentials of psychological testing. (Rev. Ed.) New York: Harper & Row, 1960.

Dahl, T. Toward an evaluative methodology for criterion-referenced measures: Objective-item congruence. Paper presented at the annual meeting of the California Educational Research Association, San Diego, April, 1971. (Also CSE Working Paper No. 15)

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1965, 22, 15-25

Garvin, A. D. The applicability of criterion-referenced measurement by content area and level. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 514-521.

Harris, M. L. & Stewart, D. M. Application of classical strategies to criterion-referenced test construction: an example. Paper presented at the annual meeting of the American Educational Research Association, New York, February, 1971.

Ivens, S. H. An investigation of item analysis, reliability, and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, August, 1970.

Klein, S. Evaluating tests in terms of the information they provide. Evaluation Comment, 1970, 2(2), 1-6.

Livingston, S. A. The reliability of criterion-referenced measures. Paper presented at the annual meeting of the American Educational Research Association, New York, February, 1971.

Meyers, J. L. Fundamentals of experimental design. Boston: Allyn & Bacon, 1966.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Popham, W. J. Indices of adequacy for criterion-referenced test items. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Popham, W. J. & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6(1), 1-9.

Price, D. B. A group approach to the analysis of individual differences in the randomness of guessing behavior on multiple choice tests and the development of scoring methods to take such differences into account. Research Bulletin, No. 64-59. Princeton: Educational Testing Service, 1964.

Skager, R. W. Student entry skills and the evaluation of instructional programs: A case study. CSE Report, No. 53. Center for the Study of Evaluation, University of California, Los Angeles, 1969.

Thompson, W. A. The problem of negative estimates of variance components. Annals of Methematical Statistics, 1962, 33, 273-289.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.